Fibronectin III9-10 adsorption to self-assembled monolayers and interdomain orientation in the context of material-driven fibrillogenesis studied with molecular dynamics simulations

Bieniek, Mateusz

Awarding institution:
King's College London

King's College London
Faculty of Natural & Mathematical Sciences
Department of Physics

Fibronectin $III_{9\text{-}10}$ Adsorption to Self-Assembled Monolayers and Interdomain Orientation in the context of Material-Driven Fibrillogenesis studied with Molecular Dynamics Simulations

Mateusz K. Bieniek

Doctoral thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physics at King's College London, October 2019

Abstract

Fibronectin (Fn) is a structural glycoprotein that plays a major role in the extracellular matrix (ECM) and communication with the intracellular environment. Fibronectin fibrillogenesis is triggered after its quaternary structure transitions from a compact to an extended conformation. *In vivo*, this takes place via the $FnIII_{9\text{-}10}$ - integrin $\alpha5\beta1$ binding after which the intracellular skeleton exerts force on the structure. However, fibrillogenesis can also take place on certain materials, such as ethyl acrylate (EA) functionalised self-assembled monolayers (SAMs). Surprisingly, fibronectin fibrillogenesis does not take place on a chemically similar surface, methyl acrylate (MA) SAMs. Classical molecular dynamic simulations are used to investigate the adsorption of $FnIII_{9\text{-}10}$ and its relation to fibrillogenesis. I show that surface water hydration explains why Fn adsorbs to EA SAMs but not to MA SAMs, which differs only by one extra methylene bridge. The $FnIII_{9\text{-}10}$ domains adsorb with the 9^{th} domain to EA SAMs with the CHARMM36 forcefield, which leads to exposure of the RGD and PHSRN motifs for potential binding. I also show, however, that the 10^{th} domain by itself adsorbs well to the surface. The two domains adsorb well to methyl SAMs in a non-specific way, always burying the two motifs in the surface. I reproduce the simulations with the CHARMM36m forcefield and show that, in contrast to CHARMM36, $FnIII_{9\text{-}10}$ adsorbs with the 10^{th} domain to EA SAMs, which also buries the motifs in the surface. Moreover, with CHARMM36m, adsorption of $FnIII^{9\text{-}10}$ to methyl SAMs converges to the same adsorption state, which always makes the RGD and PHSRN motifs unavailable for binding. Furthermore, I show that the forcefield CHARMM36 does not reproduce the correct behaviour of the $FnIII_{9\text{-}10}$ in bulk water. Unfolding of the tertiary structure of the 9^{th} domain takes place, which is inconsistent with the experimentally determined stability and structure of the domain. I further show that this problem disappears when the CHARMM36m forcefield is used, which I then use to investigate the different $FnIII_{9\text{-}10}$ interdomain orientations. I show that the two domains have a major preference for the interdomain conformation found in the crystal structure with PDB 1FNF. However, I also show that there are two less-common interdomain orientations which appear to be stable, and one of those has an acute angle between the two domains. Our investigation of $FnIII_{9\text{-}10}$ is discussed in the light of the available experimental data and any extrapolations are discussed with respect to their potential biological role.

Acknowledgements

I would like to thank the many academics supporting me through this journey. My first thanks go to my supervisor Christian D. Lorenz based at King's College London for his advice and support shown during this four-year period. Then I would like to thank William D. Taylor at the Francis Crick Institute while also wishing him a great career now that he is retired. At the latter institute, I had the luck to spend much time with the postdoctoral fellow Enrico Spiga, who was always willing to share his experience in the field of molecular dynamics as well as his sincere opinions in the times of trouble. To Paul Smith, for his friendship, support, constructive discussions and the perfect mixture of empathetic listening and criticism. To Lisa Grant, for her support and love and proofreading - part of the love, and listening to my endless complaints. To Marius Kausas for his friendship in the struggles. To my parents and siblings, for their support and for always trying to do what is best. To my past and present PhD group for their contributions, discussions and friendships, including Daniel Allen, Robert Ziolek, Sze May Yee, Jemma Trick, Jirawat Assawakhajornsak and Hrachya Ishkhanyan. To Jianan Bao for the always-present fresh bread available in the house. For the many interesting discussions and coffee breaks, thanks to Bhavin Khatri and Raphael Chaleil. I would like to thank for their support Joanna Czerwińska, Edgar Herrera-Delgado and Lorena Garcia-Perez. Then I would like to thank the team in Glasgow headed by Prof. Salmeron-Sanchez, particularly Virginia Llopis-Hernandez and Annie Zhe Cheng. To my thesis committee supervising my progress: Paul Bates, Argyris Politis and Justin Molloy for their advice and fruitful discussions. I would like to thank the IT departments for their help with the HPC cluster(s): at King's College London this includes Terrance Arter, and at the Francis Crick Institute this includes Miguel Oliveira, Joerg Sassmannshausen and John Bouquiere. Furthermore, I am grateful the the computational resources of Thomas HPC and ARCHER which is a UK National Supercomputing Service. At the end, I would like to thank Karol Kuczmarski, the author of a free programming book on C++ that has initiated my intellectual journey - I can only wonder how different my life would be if it was not for that book and the confidence it instilled in the teenager me.

Dedication

I dedicate this work to the struggle of PhD students and academics, with the looming prospects, they persevere.

'You are what you repeatedly do'

Aristotle

Contents

List of Tables

List of Figures

Introduction

Fibronectin is a large glycoprotein and an important block of the extracellular matrix. In **Chapter 1** first I introduce the topic of protein adsorption to biomaterials after which I move on to highlight the structure and the complex modular nature of fibronectin. I will discuss how these modules, or domains, are organised together to give rise to fibronectin's quaternary structure. Fibronectin comes in two forms, namely the soluble inactive compact and the insoluble extended conformations. These are discussed in the context of the process required for fibronectin to transition from the compact to the extended form. It is this transition that initiates fibronectin fibrillogenesis. I briefly present how the transition from compact to extended conformation after binding to integrin receptors occurs in an *in vivo* environment. I also talk about how fibronectin adsorbs to various materials, and how some materials (e.g. poly(ethyl acrylate)) can be used to initiate fibrillogenesis, which is a major aspect of the research in this thesis. **Chapter 2** introduces classical molecular dynamics simulations, which are the main computational tool in this thesis. I will present the theory as well as the practical challenges and the latest developments that are relevant to this work. Then I begin the presentation of my own work in **Chapter 3** where I summarise the adsorption of two domains, $FnIII_{9\text{-}10}$, which contain the PHSRN and RGD motifs, respectively. The adsorption of the two domains to ethyl-acrylate self-assembled monolayers (EA SAMs) is described, which is used to model the poly(ethyl acrylate) interface. This adsorption is compared to the behaviour of $FnIII_{9\text{-}10}$ on another surface, MA SAMs, which models the polymer poly(methyl acrylate). This polymer is different only in that it is missing one methylene bridge, but this small difference completely changes the experimentally observed adsorbed structure of fibronectin. The chapter concludes showing that the difference in adsorption to MA and EA SAMs is due to the hydration of the

surfaces. During the adsorption of $FnIII_{9\text{-}10}$ to EA SAMs it has been found that the 9^{th} domain drives the adsorption while weakening the adsorption of the 10^{th} domain, which nevertheless continues trying to adsorb. And this is the topic of **Chapter 4** which shows that the 10^{th} domain adsorbs well to surface in the absence of the 9^{th} domain. In **Chapter 5** I present the nature of the adsorption of the same two fibronectin domains to hydrophobic Methyl SAMs. I show that both of the $FnIII_{9\text{-}10}$ domains adsorb to the surface quickly and well but non-specifically. I show that the RGD and PHSRN motifs are buried in the surface, making them unlikely to be available for integrin binding. Moreover, I show that the 9^{th} and 10^{th} domains rotate with respect each other which led me to question how much the 10^{th} domain can rotate and if there is a conformation in which the two domains can adsorb to the EA SAMs at the same time. Therefore focus is shifted to the interdomain orientation of the two domains, which led to the finding that the CHARMM36 force field is overestimating the protein-protein interactions, and thus leads to the partial break down of the 9^{th} domain's tertiary structure, which is discussed in detail in **Chapter 6**. This work is followed by an investigation utilising the more recent forcefield CHARMM36m while sampling the interdomain interface in **Chapter 7**. The problem with overestimated domain-domain interactions disappears with the newer forcefield CHARMM36m. By simulating the 10^{th} domain at different orientations with respect to the 9^{th} domain, I show that the major domain-domain conformation is the same as the original crystal structure (PDB:1FNF). However, the two domains are capable of assuming different orientations with respect to each other, presenting two additional conformations. **In Chapter 8**, with the new force field, I confirm my previous results. I show how analysis of adsorption to surfaces can be more easily described with the use of clustering (DBSCAN) and discuss the similarities and differences found during the adsorption in the Chapters 3, 4 and 5. After that, this thesis is concluded in **Chapter 9** in which all of the results are discussed, and potential future directions for the research summarised in this thesis are described.

Motivation and Objectives

The motivation for this thesis was to use atomistic molecular dynamics simulations to study the adsorption of fibronectin $III_{9\text{-}10}$ domains. In doing so, I wished to provide a detail description of the molecular mechanisms that govern the material-driven fibrillogenesis of fibronectin that is of importance to biomaterial research.

Contributions

In addition the various scientific contributions summarised in Chapter 3 - 8 I have contributed to the open source software MDAnalysis and PyMOL during my PhD. In MDAnalysis, together with Paul Smith I improved the *waterdynamics* package by fixing bugs and improving the implementation of discrete autocorrelation and intermittency. Intermittency helps users to treat systematically fluctuations around the rigid selection boundaries in MDAnalysis. In collaboration with Paul Smith, I was awarded the Warren L. DeLano Memorial PyMOL Open-Source Fellowship. With this funding, we made useful additions to the PyMOL software package which allows for it to work with molecular dynamics trajectories without loading them into memory, and also we have integrated new tools to generate interactive analysis/plots and then to easily include them within a latex document.

These contributions, together with the code for MDAnalysis, have been described in Appendix A. Furthermore, the analysis code written during this thesis is documented and described in the Appendix as well.

Statement of Originality

I hereby declare that this work was done by me only.

Publications

The work which I summarised in Chapter 3 of this thesis has been accepted for the following publication:

Mateusz K. Bieniek, Virginia Llopis-Hernandez, Katie Douglas, Manuel Salmeron-Sanchez, Christian D Lorenz (2019) 'Minor Chemistry Changes Alter Surface Hydration to Control Fibronectin Adsorption and Assembly into Nanofibrils', Advanced Theory and Simulations. John Wiley & Sons, Ltd, p. 1900169. doi: 10.1002/adts.201900169.

Chapter 1

Fibronectin

Fibronectin (Fn) is a large glycoprotein that is a major component of the extracellular matrix. It is made up of many modules, most of which are joined by very short linkers. These modules, or domains, bind to many biomolecules and have different stabilities. In addition, their overall number depends on the splicing. The domains interact together to form a *compact* complex quaternary conformation. In order to initiate fibrillogenesis fibronectin has to transition from this compact conformation to an extended one. The extended conformation exposes previously unavailable sites allowing fibronectin to bind to itself and then to embed itself into the extracellular matrix. Fibronectin fibrillogenesis creates a biologically active environment that is sensed and acted upon by the cells. *In vivo*, fibrillogenesis begins when fibronectin is pulled by integrin receptors. However, in this thesis I focus on an alternative way to initiate fibrillogenesis, which is materials-driven fibrillogenesis. In this approach, a material interface causes fibronectin to adapt an extended conformation upon adsorbing to the interface and therefore leads to the process of fibrillogenesis. In this chapter I introduce the topic of protein adsorption to biomaterials. I then describe fibronectin in more detail, from its biological context, its structure, its adsorption to different materials, and its ability to form biologically active networks *in vivo* and on different materials.

1.1 Protein Adsorption & Biomaterials

Adsorption of proteins is an important and complex process with potential ramifications for an array of therapeutics. The closest field concerned with it is that of biomaterials, materials which have been designed for *in vivo* use. Understanding how a human organism interacts with biomaterials, with emphasis put on protein-surface interactions, is crucial to furthering the use of materials in applications such as bone replacement, artificial pacemakers, cochlear implant, tissue engineering [4], wound healing [5], and the use of stents [6, 7]. Similarly, understanding protein adsorption to non-artificially introduced surfaces, such as teeth and bones, which primarily constitute hydroxyapatite, can help to define the required properties to design new biomaterials [8, 9]. And there is a lot of scope for improvement, as the use of current implants can lead to thrombosis formation, toxicity, inflammation, infection [10, 11], fibrous encapsulation, immunogenic response and even rejection [12, 13, 7]. These problems are far from resolved as became clear when the International Consortium of Investigate Journalists (ICIJ) released the "Implant Files", showing that many inadequately tested medical devices and implants were used without a sufficient regulatory oversight in Europe and North America [14]. This was not, however, completely unforeseen, as the Royal College and the British Orthopaedic Association (BOA) warned the public of lax regulations back in 2012 [15]. Correspondingly, due to the challenges involved, the projections of future revision of hip and knee implant arthroplasty indicate that their number will only increase [16]. Besides the direct implementation of biomaterials in therapeutics, understanding protein adsorption can contribute to the field of stem cell research, where the adsorbed protein-layer affect cell growth and differentiation [17, 18].

Numerous proteins are known for their adsorption to surfaces. Abundant (1 mg/ mL) plasma proteins like Albumin [19, 20], Transferrin, Fibrinogen [21, 6] and Immunoglobulins have been studied in the context of adsorption. Another large group includes proteins constituting the ECM, which often serves as the primary contact for the cells [22, 23] *in vivo*. This is due to the ability of ECM proteins to bind to the membrane receptors such as integrins or cadherins. In addition, ECM proteins have good adhesive properties and therefore make up the protein mix that surrounds the implanted devices [13, 24]. ECM proteins include, among others, collagen,

the most abundant protein in human body [25], laminin which is a component of the basal lamina, or protein network foundation for most cells and organs, as well as fibronectin. Proteins in contact with surfaces are often difficult to study due to the their variety, aggregation, structural changes and/or denaturation, long-term displacement, or the interactions with the immune system. Although the challenges are numerous, the potential benefits are just as numerous. A well designed protein layer can aid an implant in avoiding the undesired response of patient's adaptive immune system, which otherwise could lead to the use of immunosuppressive drugs [12].

The type of surface dictates to a large extent which (and how) proteins adsorb. Materials are characterised in terms of the desired properties which among others include mechanical properties, longevity [26], toxicity or infection-resistance [11]. The meaning of biocompatibility has been standardised by the International Organization for Standardization (ISO), which depending on the nature of body contact defines criteria such as carcinogenicity, degradation or chronic toxicity (ISO 10993). Some of the researched materials were inspired by the natural surfaces found in the human body, such as bones or enamel in teeth, whose prime constituent is hydroxyapatite. Hydroxyapatite-based or -coated materials carry a promise in areas like hip replacements or dental implants [27, 20]. However, most biomaterials in use are artificial, including metallic, polymeric and ceramic materials [28]. One successful example of a metallic material in the field of orthopaedic implants is titanium [29, 30, 7] due to its mechanical suitability and biological inertness. Other used metallic materials include gold, cobalt chromium, tantalum and 316L stainless steel [31, 32]. However, metallic surfaces struggle with corrosion and in some cases, toxicity. The widely adapted titanium alloys release aluminium and vanadium which might be linked to Alzheimer disease, neuropathy and osteomalacia [33]. Similarly, the stainless steel 316L is known to corrode and release toxic chemicals [29, 34, 35]. Another used material type are polymers, which due to their unique ability to form three dimensional complex matrices, can prove particularly useful in the field of tissue engineering [4]. Furthermore, polymers can be biodegradable, which is of great potential to drug delivery for controlling the rate at which the drug is released [4]. The different materials are often combined, as in the case of stainless steel coated with a polymer in order to create the drug eluting stents (DES)

which were found to reduce restenosis [36]. The third material type is ceramics which, when it comes to implants, have the disadvantage of poor mechanical properties [28].

The surfaces of materials are often modified by changing the topography, or by coating the surface with proteins or polymers to increase the desired physicochemical interactions [4, 37]. The materials coated with proteins emulate the properties of the native ECM and are therefore referred to as biomimetic materials [8]. These biomimetic materials attempt to use cell receptors available to aid the formation of tissues to increase biocompatibility of the material [22, 38].

1.2 Biology & Structure

Fibronectin is a major component of the extracellular matrix (ECM). It is secreted by cells in cellular environments. However, when it is secreted in the liver by hepatocyte cells it often finds its way into plasma [39]. When it is secreted in a local cellular environment, fibronectin is directly embedded into the extracellular matrix and is said to be *insoluble*. When it is secreted by hepatocytes into plasma, it becomes a major component of plasma at the concentrations of 300 - 600 μg/ml [40], and is referred to as *soluble* fibronectin. This soluble fibronectin later is found to be a component of tissues around the body [39, 41].

Fibronectin has numerous binding sites for a variety of biomolecules including integrins, collagen/gelatin, heparin, fibrin, bacteria, growth factors [42, 43], cytokines, and crucially for fibronectin fibril formation, other fibronectin molecules [44, 45]. Fibronectin is necessary during the development. The total removal of fibronectin in mice leads to a death at the stage of an embryo [46]. Its structural role in forming and maintaining the extracellular environment makes it necessary for cellular processes such as cell organisation, growth [47] and proliferation [48].

Fibronectin has been shown to play an important role in various diseases. For example, fibrils have been implicated in rare cases of glomerulopathy where large fibril deposits caused renal failure [49, 50, 51, 52]. It also affects cancer progression, and for this reason it has been targeted in cancer imaging [53, 54]. It reduces the inhibition of cell cycle progression during

lung tumour metastasis. Fibronectin can also modulate chronic inflammation, which is also a factor in cancer. Inflammation, in turn, can encourage further fibronectin and collagen deposition creating a positive-feedback loop that leads to a fibronectin-rich ECM (see review [54]).

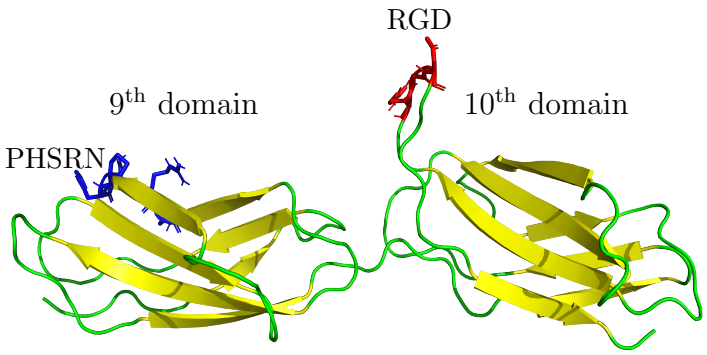


Figure 1.1: A cartoon representation of the two domains $FnIII_{9\text{-}10}$ from the crystal structure with PDB:1FNF[1]. The two domains span the residues 1327-1509. The red and blue licorice represent the RGD and PHSRN motifs, respectively. The β-sheets are coloured yellow.

Fibronectin is a dimer of polypeptide chains linked via a pair of disulphide bonds at their carbonyl terminals [55]. Each of the polypeptide chains can be thought of as beads-on-a-string, or as a list of linked domains which have three distinctive types. These domain types are labelled with Roman numerals I-III. A single polypeptide contains 12 type I, only 2 type II, and 15-17 type III domains, depending on the slicing. There are at least 20 variants of human fibronectin. The soluble fibronectin almost always have 15 type III domains [56], whereas insoluble fibronectin is more heterogeneous and often includes extra type III domains (called ADA and EDB) [44].

The domains in a dimer interact with each other to form a globular-like quaternary structure, or compact fibronectin. It is this soluble or compact conformation that is found in plasma where fibronectin-fibronectin interactions would be deadly. Fibronectin can also assume an extended conformation which has almost double the radius of gyration (17.5±0.8 nm vs 10.7±0.9 nm) as measured with light scattering [57]. The compact conformation is kept stable due to interactions between FnI_4 and $FnIII_3$ of the same polypeptide, and $FnIII_{2\text{-}3}$ and $FnIII_{12\text{-}14}$ of the different polypeptides within the same dimer [58]. These interactions are believed to be of an electrostatic nature [59]. For a review of the quaternary structure, see [60, 61, 62].

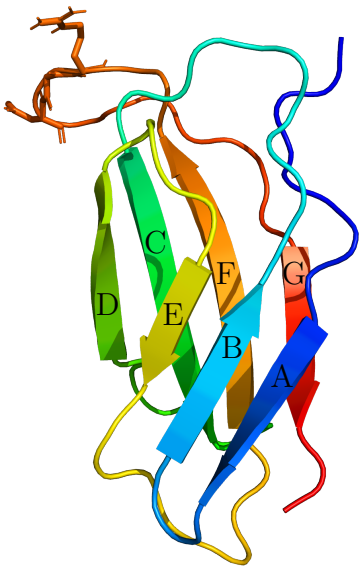


Figure 1.2: A cartoon representation of a $FnIII_{10}$ domain representing the β-sandwich structure with each β-strand labelled A to G in the direction of the carboxyl-terminus (C-terminal). The licorice represents the RGD motif. The β-sheets have rainbow colours starting with blue at the N-terminal to red at the C-terminal.

All fibronectin domain types have the same tertiary structure. Specifically, a β-sandwich structure with two opposing antiparallel β-sheets. However, the type III domains stabilise the tertiary structure of the two β-sheets with a hydrophobic core - whereas type II and I domains use intra-chain disulfide bonds for this purpose. The secondary structures in the domains remain stable even at high temperatures [63]. The type III domain has 7 β-strands in total where each of the strands is labelled with a letter from A to G from N-terminus to C-terminus (Figure 1.2). Each of the type III domains contains around 90 amino acids [64].

The two domains $FnIII_{9\text{-}10}$ are particularly known for their role in fibrillogenesis *in vivo*. Fibronectin famously binds to the integrin receptors via the RGD (Arg-Gly-Asp) motif which resides on the $FnIII_{10}$ domain [65]. It works in synergy with the PHSRN motif on the $FnIII_9$ domain, which for this reason is referred to as the "synergy region" as it enhances the binding to the $\alpha5\beta1$ [66, 67] integrin. These two domains are visualised in Figure 1.1 and they are the focus of this thesis. The RGD motif resides on the last loop of the 10^{th} domain between the β-strands F-G, whereas the PHSRN motif resides between the β-strands D-E (Figure 1.2). The motifs reside on the same face of the protein fragment which is believed to be important for PHSRN to enhance binding to integrins. However, RGD, is the more important motif respon-

sible for most of the interactions with integrins [66], and here it is seen protruding away from the two domains, which likely increases its binding availability [1].

In vivo, in order to start fibrillogenesis, the binding of the RGD motif to the integrin receptors must take place [68]. After binding, actin cytoskeleton reorganisation exerts force on the compact fibronectin dimer in order to extend it [69, 70]. The extension of the fibronectin takes place at the same time as integrin clustering which brings fibronectin molecules within sufficient proximity for self-association [71]. For a review of a cell-driven fibrillogenesis see [61]. Furthermore, fibronectin can start fibrillogenesis under other non-physiological conditions, including increased temperature, changing pH, adding salts or applying force (see reviews [72, 73]).

Once the fibronectin fibrils are established, they affect, among other things, the development processes, tissue homeostasis, fibrosis and cancer progression. Furthermore, the growing field of mechanobiology is improving our understanding of how integrins sense the rigidity of the environment through focal adhesions. Cells act differently depending on the properties they sense, such as rigidity, which in turn depend on whether the environment is healthy or strained [74]. For example, some of the cryptic sites on fibronectin fibers become available after being stretched [75, 76], while other bindings are destroyed in the process. Therefore fibrillogenesis is important for many cellular processes.

1.2.1 Fibronectin Adsorption

Fibronectin with its mosaic nature comprising hydrophobic and hydrophilic components adsorbs well to a variety of surfaces. It absorbs particularly well to hydrophobic surfaces such as gold [77, 78], with 35% more surface mass density on gold than on tantalum oxide and titanium oxide [77]. In the presence of gold nanoparticles, fibronectin appears to unfold on untreated gold nanoparticles [78]. Interestingly, fibronectin adsorbed to titanium dioxide forms globular aggregates, rather than extended fibrills [79]. Another group compared the behaviour of fibronectin on anatase and rutile titanium dioxide, finding a better osteogenic activity on the anatase film, which the group attributed to the larger number of hydroxyl groups [80]. Fi-

bronectin adsorbs well to hydroxyapatite nanoparticles, but the mode of adsorption depends on the size of the nanoparticles. The larger particles (which have a weaker negative potential) lead to an extended fibronectin conformation [27]. On bactericidal surfaces such as nanostructured black silicon surfaces, which is full of "spikes", fibronectin adsorbs across them in a convex manner [81]. Similarly, the surface roughness of poly-pyrrole doped with dextran sulfate has no effect on fibronectin adsorption [82].

Interestingly, an increased positive potential due to oxidation led to more extended fibronectin conformation [82]. Another confounding effect in the adsorption of fibronectin is the presence of other proteins [78]. Fibronectin's co-adsorption with human serum albumin affected its ability to expose cell binding region [83]. Gold nanoparticles which were pre-coated with other proteins showed that fibronectin interacts with the proteins rather than displacing them [78]. However, it should be noted that fibronectin adsorption can be affected by the amount of time fibronectin is given to settle down on the surface, as well as the concentration [84].

In order to understand the effects of the surface chemistry on the adsorption of fibronectin, the model self-assembled monolayers are often used. Typically these are alkanethiols on gold terminated with different functional groups. On the hydroxyl-group terminated surface, fibronectin exposes more of the cell-binding region [85], which might explain why more osteogenic activity was observed on the anatase titanium dioxide [80]. However, the cell-binding region was found to be less available on hydrophobic surface [85], on which fibronectin adsorbs very well [86]. This shows further that the adsorption is necessary but not sufficient, as the availability of the cell-binding region determines cell adhesion. This was further confirmed when it was found that fibronectin had a tighter binding but underwent denaturation on hydrophobic surfaces [86].

1.2.2 Modelling Adsorption

Adsorption can be described as interactions between the atoms in a protein and atoms of a given surface. Therefore, the properties of atoms can be used to aid our understanding of adsorption. Assuming that no chemical interactions take place, these interactions can be summarised with three components: electrostatic, van der Waals (dispersion) and hydrophobicity (entropic inter-

actions) [87]. Electrostatic interactions are captured by Coulomb's Law and function over the longest distance, which can be measured with the Debye length that takes into account the ionic strength of the medium [87]. Electrostatic interactions, depending on the protein and surface charges, ion-screening and the dipole of the protein affect how the the protein approaches the surface [88, 89]. The dispersion interactions are often weakly attractive, but their cumulative effect can change the orientation in which the protein is adsorbed to the surface. Hydrophobicity is described as the entropic effect of water. As the hydrophobicity often stabilises the tertiary structure of a protein, adsorption can lead to an increase in conformational entropy, or loss of structure [90, 87].

Computational modelling approximates these interactions to study protein adsorption which can aid the interpretation of experiments. The different approaches include the Monte Carlo simulations, atomistic molecular dynamics simulations which I introduce in Chapter 2, and enhanced sampling methods which help estimate the free energy landscape of a system [91]. Computational modelling generates highly detailed molecular-level data. Such a detail can be difficult or impossible to obtain experimentally, which often relies on indirect measures. Examples include the use antibodies to probe which face of the protein is available for binding, and global measures such as circular dichroism or other spectroscopy-based approaches.

Due to its large size, the entire fibronectin cannot be feasibly studied with atomistic simulations. However, due to the modular nature of fibronectin, it is possible to study its components. The most frequently modelled fibronectin modules are the biologically active cell-binding $FnIII_{9\text{-}10}$. In this paragraph I summarise the available computational modelling studies. The fibronectin cell-binding region adsorbs to the rutile titanium with a stable orientation [92]. On charged surfaces, the adsorption of the same area, $FnIII_{8\text{-}10}$, was found to be driven mainly by electrostatics [93]. However, the strongly surface-bound water hindered the adsorption [93]. On polar and hydrophobic surfaces the adsorption was found to be non-specific [93]. Consistently, the smaller fibronectin part, $FnIII_9$, adsorbed rapidly on the silica surface driven by the electrostatics, and non-specifically to the hydrophobic gold surface [94]. Although $FnIII_9$ has -1e charge, it adsorbs to negatively charged mica [94]. Another group used Monte Carlo and Molecular Dynamics simulations to study $FnIII_{9\text{-}10}$ and $FnIII_9$ showing that the hydrophobic surface

was deactivating the RGD motif, whereas hydrophilic surfaces did not suffer from this problem [95]. Crucially, they also found that the more positive surface exposed the RGD motif for binding [95]. This is in agreement with an experiment which showed that fibronectin had an extended conformation on a similarly-charged surface [82]. The same area, $FnIII_{10}$ and $FnIII_{7\text{-}10}$, was studied on hydroxyapatite surfaces using parallel tempering Monte Carlo and Molecular Dynamics [96]. They also showed that positively charged surface improved adsorption and the accessibility of the RGD motif, and added that surface artefacts have the ability to trap guanidine groups [96]. Artefacts in rutile titanium also strongly affected the orientation of the cell-binding region during adsorption [92].

1.2.3 Material Driven Fibrillogenesis

Fibronectin can form biologically active fibrils without the involvement of cells or integrins. A recently discovered example of a surface which can do this is the polymer poly(ethyl acrylate), or PEA [97]. On this polymer fibronectin forms networks which, among other things, improve fibroblast focal adhesion development, actin filament maturations and myogenic differentiation [98]. Similar fibrils were also observed on a biocompatible polymer poly(methyl methacrylate), which has been previously used in implants [99]. While fibronectin adsorbs to many other surfaces, fibronectin networks are rarely observed. For example on mica surfaces [99] a single fibronectin molecule has very similar dimensions and shape to that of a single fibronectin molecule on a poly(ethyl acrylate) surface [97]. Specifically, around 90 nm long, which classifies it as a compact conformation as the radius of gyration for compact fibronectin is 10.7 ± 0.9 nm [57, 74]. Despite these similarities, fibronectin networks are not formed on the mica surface. Furthermore, in both cases the fibronectin size indicates that the molecule is not fully extended, which is not consistent with the hypothesis that fibronectin extends itself on the surface.

Another polyalkyl acrylate polymer similar to PEA is poly(methyl acrylate) - it has only one fewer methylene bridge in its side chain. However, this is enough to stop fibrillogenesis [98]. The length of the side chain of the polyalkyl acrylate family, to which they both belong, has been further investigated showing that longer, more mobile and more hydrophobic side chains also

lead to fibronectin network formation. Interestingly, longer chains were also associated with the increased exposure of the integrin binding domains [100]. However, hydrophobicity is not the only factor. A surface that is too hydrophobic decreases fibronectin's biological activities. A superhydrophobic polystyrene surface was compared to smooth polystyrene and fibronectin showed less adsorption and fewer cells formed mature focal adhesions [101].

Fibronectin forms networks on electroactive poly(vinlyidene fluoride) films (PVDF), specifically the non-poled β-PVDF at the concentration of 2 μg mL^{-1} [102]. Across the different PVDFs tried this non-poled PVDF was the most hydrophobic (almost 80° contact angle). Strangely, the integrin blocking monoclonal antibody HFN7.1 at fibronectin concentration 5 μg mL^{-1} showed that this site is 3 times more available on β negative poled surface than on the hydrophobic β-non poled PVDF. Despite the formation of fibronectin networks, the integrin site is less available. Furthermore, the cell numbers were higher on β poled -/+ than on the non-poled β PVDF. This means that fibronectin networks can present different domain epitopes on different substrates.

The behaviour of fibronectin on hydrophobic and hydrophilic surfaces was further investigated with atomic force microscopy in [103]. With an even smaller concentration of the protein (1 μg ml^{-1}) silica, mica and hydrophobic surfaces were probed. On the hydrophobic surface fibronectin was found to be in the smaller, or compact, form. In both silica and mica the protein was classified as extended, with the length estimated to be 123$\pm$28 nm and 121$\pm$25 nm. These look very similar to the single molecules found on poly(ethyl acrylate). It was suggested that the hydrophilic surface disturbs the electrostatic interactions that keep fibronectin in the compact form. However, despite the similarities between the cases, it was shown before that at higher concentrations fibronectin did not form fibril networks on mica [99]. In other words, the extension of fibronectin, or some form of it, is necessary, but not sufficient for fibrillogenesis.

Surface rigidity is an important factor in fibronectin adsorption. In [104], ten self-assembled monolayers were tested, each with a different functional group that was either highly hydrophobic, charged or polar. Fibronectin adsorbed to all of them. However, this adsorption did not translate into the same biological activity on every surface. F-acting stress fiber reorganisa-

tion or neurites growth was quantified on each of the surfaces. F-acting reorganisation was observed to a larger extent on -SiOH and -Br, and less so on -CH_3 and C=C. However, neurite formation was often poor when F-acting reorganisation was observed. Furthermore, blocking fibronectin-integrin binding with synthetic RGD peptides affected F-acting reorganisation without a consistent effect. Similarly, the most likely explanation is that different domain epitopes are presented depending on the terminal functional group in the self-assembled monolayer which affects F-acing reorganisation and neurite formation.

The entire fibronectin is not necessary in order to change cellular fate despite fibronectin fragments often being unable to form fibrils. The adsorption of the fibronectin $FnIII_{7\text{-}10}$ fragment to biomaterials can help with the integration of the material with the human body. For example, coating stainless steel screws with the four domains enhanced bone-screw fixations in both healthy and osteoporotic rats. The integrin $\alpha_5\beta_1$ was a crucial part in enhancing the bone-screw fixations [105]. However, this approach has shortcomings because it excludes potentially important parts of the glycoprotein. For example, in tissue healing, fibronectin networks on the polymer poly(ethyl acrylate) were used to help deliver growth factor, whose injection is often toxic to local environment [106]. In that study, the fibronectin matrix presented the bound bone morphogenetic protein 2 (BMP-2) which improved regeneration of non-healing bone defects.

Fibronectin fibrillogenesis is a complex process that is closely linked to the adsorption of the protein. Understanding how it adsorbs to surfaces and how that is related to biological function is therefore the topic of this thesis. In the following chapter I introduce classical molecular dynamics simulations which I used to undertake this study.

Chapter 2

Molecular Dynamics

One of the earliest historical use of modelling in biology goes back to the publication "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid" by Watson & Crick. With experimental clues, including the infamous use of Franklin Rosalind's X-ray pictures, James Watson and Francis Crick devised a structural model of DNA. In their two-page publication, the language of life was revealed. Since then the field of modelling has grown exponentially. Furthermore, with the explosion in the number of available X-ray crystal structures, the modelling of the structure-function relationship became the cornerstone of biology. Classical atomistic molecular dynamics simulations offer great potential to modelling molecular phenomena. By offering insights into atomistic dynamics, this computational approach can reveal biological properties that might not be attainable experimentally. The significance of the field was recognized in 2013 when the Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Arieh Warshel for the development of multiscale models for complex chemical systems. In this chapter I introduce molecular dynamics simulations and the models (forcefields) used to study proteins.

2.1 History

In 1953, Francis Crick and James Watson published a new model of DNA [107]. Their ideas were based on experimental evidence including the ratio of base pairs (Chargaff's rules) as well as the X-ray structure from Rosalind Franklin which formed the basis for the discovery. However, less well known is the influential work done with aluminium models which they used to try different arrangements of the nucleotides. Since then, modelling has evolved significantly with one of the successors being the field of molecular dynamics. Molecular dynamics is a computational approach to modelling that solves Newton's equations of motion for a system of particles in order to study their dynamic properties.

One of the earliest simulations goes back to 1959 where several hundred interacting particles were simulated to study the liquid phase transition [108]. In these simulations hard spheres were used to approximate the behaviour of atoms. Nowadays, the interactions between particles are governed by interatomic potentials, which non-linearly account for their charges, dipole moments and structural properties. Since the study in 1959, the field of molecular dynamics has evolved significantly. Large improvements have been seen in the models (their functional form and parametrisation), hardware approaches, sampling techniques and analysis tools.

In the last decade molecular dynamics simulations have been increasingly used as a complementary tool to experiments. In addition to this, the increasingly refined models, or forcefields, created for molecular dynamics found their use in many other fields. For example, the pharmaceutical industry uses molecular dynamics and its forcefields for high-throughput scanning of small molecules [109]. The forcefields have been adapted in the world of bioinformatics and in the different sampling approaches focused on protein-protein interactions. A great illustration of this is found in the 2012 round of Critical Appraisal Skills Programme (CASP) which is the golden structure prediction benchmark. CASP is a double blind experiment that determines the progress in computational protein structure prediction. It does so by inviting computationally predicted structure submissions while the structures are being resolved experimentally. The use of the molecular dynamics forcefields has been highlighted in CASP12: "eight out of the top ten CASP12 refinement methods use MD with recently developed physics-based potentials

... while the remaining two use MD with a hybrid knowledge-based/physics-based potential" [110]. However, the presence of molecular dynamics can also be seen in many other areas. Two of these areas come with unique difficulties that the field can help with: the intrinsically disordered proteins [111] and the membrane proteins [112].

2.2 Background

Classical molecular dynamics is a physics based simulation approach that relies on the Born-Oppenheimer approximation. It states that the motion of atomic nuclei and electrons are largely independent, and the electron cloud will instantly adjust to a new position of the atomic nucleus. For this reason, the electrons are not treated explicitly, and are incorporated into atoms which are treated as beads obeying Newton's laws. Another important axiom in the field is the ergodic hypothesis which states that, given an infinite amount of sampling time, all microstates have an equal probability of occurring. This means that with sufficient sampling the system will explore the entire phase space, regardless of the starting system configuration. With this property, the ensemble averages can be calculated as meaningful values that represent the more likely states of the phase space.

In order to carry out a simulation an initial system has to be constructed. In a typical case a system includes a protein structure, water atoms and counterions. The protein structure is often an X-ray crystallography structure, taken from the protein data bank [113]. After the system is constructed, initial velocities need to be assigned to the atoms in the system. These velocities are sampled from the Maxwell-Boltzmann distribution at a desired temperature. Once the velocities are assigned, the simulation begins by repeatedly solving Newton's equation of motion:

$$F = ma$$

Where F represents force, m mass and a acceleration and states that the net force applied to a body results in a proportional acceleration. The equation can be rewritten as a derivative of

the potential energy of the system, U, with respect to the atomic positions, r:

$$\frac{-dU}{dr} = m\frac{d^2r}{dt^2}$$

Where r represents the coordinates of the atoms and t describes times. The potential energy is the sum of bonded and non-bonded forces:

$$U = U_{bonded} + U_{nonbonded}$$

The bonded potential energy terms capture the intramolecular behaviour and structure of the biomolecule (See Figure 2.1, left and middle). Using these potential energy terms, for example, the atoms of a functional group can be "restricted" in movement in a way that mimics its real behaviour. For example, when two bonded atoms move away from each other, a restoring force brings the two atoms closer to their preferred configuration. The situation is the same for an angle. The latter two terms, dihedral and improper dihedral angles, ensure that atoms stay in the right plane(s). The bonded potential energy terms include:

$$U_{bonded} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2$$
$$+ \sum_{dihedrals} K_\chi(1 + cos(n\chi - \delta)) + \sum_{impropers} K_{imp}(\varphi - \varphi_0)^2$$

b_o, θ_0 and φ_0 are the equilibrium values around which the harmonic potentials oscillate. For the dihedral term, cosine function is used to allow for multiple different equilibrium configurations. The symbol δ describes the dihedral angle, whereas $n\chi$ determines the number of equilibrium points. The K constants determine the steepness of the harmonic potential. The non-bonded potential energy terms approximate the intermolecular forces which include the Coulomb forces and the van der Waals which is modelled with the Lennard-Jones potential [114]:

$$U_{nonbonded} = \sum_{nonbonded} \left(\epsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^6)] + \frac{q_i q_j}{r}\right)$$

Where the ϵ defines the depth of the potential well, r is the distance between the two atoms, σ is a distance at which the Lennard Jones potential is at its minimum, and $q_i q_j$ define the electrostatic charges of the two atoms. The set of potential energy terms is referred to as a forcefield and each of the terms is visualised in Figure 2.1. Most of the biomolecular forcefields in use follow the same functional form [115, 116]. These forcefields are additive, which means that the potential energy terms from each of the interaction-pairs are added together for each atom.

The most well known families of biological additive forcefields include GROningen MOlecular Simulation (GROMOS) [117], Assisted Model Building and Energy Refinement (AMBER) [118], Optimized Potentials for Liquid Simulations (OPLS) [119], and Chemistry at HARvard Molecular Mechanics (CHARMM) [120]. Particularly of interest in this thesis, these include full refined sets of parameters for proteins and nucleic acids.

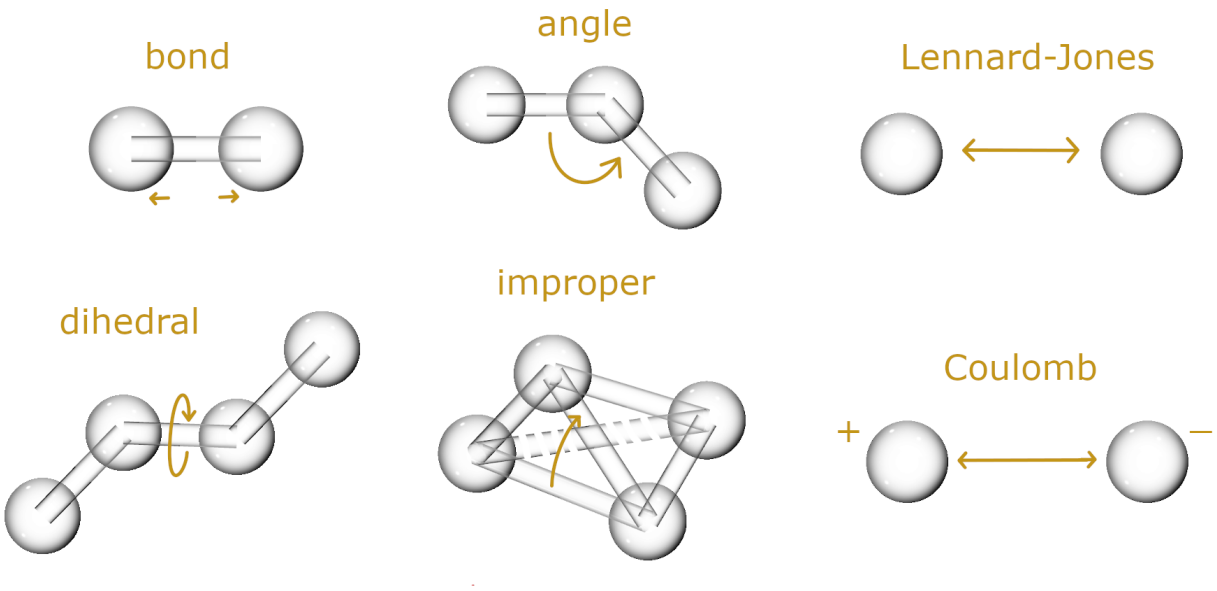


Figure 2.1: The visualised potential energy terms in a typical forcefield. The intramolecular terms include the bond, angle, dihedral and improper angle whereas the intermolecular interactions are described using only two terms. The van der Waals force is modelled using the Lennard-Jones potential, where the electrostatic interactions are modelled with the Coulomb term.

The potential energy is used to update the atomic velocities, which after a short fraction of time leads to a new position, which in turn requires recalculating the potential energy. This iterative process continues as long as necessary to sample the relevant states. This integration

process can be carried out with the leapfrog algorithm that is a fast symplectic integrator which means that it conserves the energy of the particles. This integrator is used throughout this thesis. It has the following form:

$$a_i = F(x_i)$$
$$v_{i+1/2} = v_{i-1/2} + a_i \Delta t$$
$$x_{i+1} = x_i + v_{i+1/2} \Delta t$$

where x_i is position at step i, $v_{i+1/2}$ is the velocity at step $i + 1/2$, and $a_i = F(x_i)$ is the acceleration at step i, whereas Δt is the size of each time step. Integrating the equations of motions with the leapfrog algorithm results in the microcanonical ensemble (NVE). This means that the number of particles (N), volume (V) and the total energy of the system (E) are conserved. However, this ensemble does not capture *in vivo* or *in vitro* behaviours, where the temperature and pressure are maintained. The canonical ensemble (NVT) offers an improvement by keeping a constant temperature (T) whereas constant pressure is maintained in the isothermal-isobaric ensemble (NPT) maintaining a constant pressure (P).

Different algorithms can control temperature and pressure and they are referred to as thermostats and barostats, respectively. In general, thermostats have to fix the average temperature of the system while allowing for realistic fluctuations in the system to occur. The thermostats used in this thesis are Berendsen [121], stochastic velocity rescaling (also known as stochastic Berendsen) [122] and Nosé-Hoover [123]. With a Berendsen thermostat, the kinetics of the system is corrected according to $\frac{dT}{dt} = \frac{T_0 - T}{\tau}$ where T_0 is the reference temperature. This means that a temperature deviation decays exponentially with a time constant τ. A Berendsen thermostat, however, suppresses the fluctuations of the kinetic energy, and consequently is unable to reproduce the canonical ensemble. For this reason a Berendsen thermostat is used mostly to equilibrate the system, but rarely in production. Stochastic velocity rescaling is the Berendsen thermostat with the added stochastic term that aids it in generating the correct kinetic distribution [122]. The Nosé-Hoover in turn adds extra degrees of freedom to the particles which act as a thermal reservoir [124, 123]. A Nosé-Hoover thermostat adds a friction force that is

proportional to each particle's velocity and a friction parameter ζ. This friction parameter is a fully dynamic quantity with its own momentum and equation of motion. Nosé-Hoover generates trajectories consistent with a canonical ensemble. All of the described thermostats use a non-fluctuating time step and allow for separating the systems into groups whose temperature can be maintained separately. This separation is useful for protein and liquid systems to ensure that both have a correct average temperature. Another useful property of thermostats is that they help avoid energy drifts which are caused by the accumulation of numerical errors during the integration.

A barostat maintains the pressure of a system at a desired value. In the most basic case it scales the size of the system box in every direction, which either leads to relaxation of atom-distances, or to their compression. The Berendsen barostat uses this approach. However, it does not generate the true NPT ensemble. This is in contrast to the Parrinello-Rahman barostat [125, 126], which generates a correct NPT ensemble. The Parrinello-Rahman barostat is anisotropic and is based on the Nosé-Hoover approach. It contains an extra term that resembles the friction term in the Nosé-Hoover thermostat. Furthermore, Parrinello-Rahman typically requires more time in order to achieve equilibration, which is why it is often used only in the production stage. This anisotropic property means that different pressures can be used for the different spatial dimensions. This is an important quality when simulating systems with membranes or surfaces, where the pressure is not the same in every dimension.

In simulations, the integration timestep is very small. The fast oscillating hydrogen-bonds require a 0.5 femtosecond timestep. These oscillations, however, can be neglected when the hydrogen-bonds vibrations are not studied or relevant. Ignoring the hydrogen-bonds means that the timestep can be increased up to 2 femtoseconds, but in order to ensure the stability of the system, a constraint algorithm such as SHAKE [127] or LINCS [128], which ensure that the distance between atoms is correctly maintained, have to be used. The algorithm corrects the hydrogen-bond oscillations and ensures that the distances abide by the constraints, allowing for significant improvement in the simulation sampling by up to four times.

Another optimisation concerns the computing performance and is related to the non-bonded

potential terms. The Lennard-Jones quickly becomes negligible with distance and is therefore often made to go to zero continuously after around 12 Å. However, the electrostatic interactions have to be calculated from one atom to every other in the system. A naive implementation therefore has the computational complexity of $O(n^2)$. However, the Particle Mesh Ewald summation (PME) algorithm can be used to decrease the computational complexity to $O(n \log n)$. PME divides the calculations into short-range and long-range contributions. The short range is calculated using Coulomb's Law, whereas the long-range contributions are evaluated in Fourier space.

Due to the performance limitations, molecular dynamics simulations are only capable of handling relatively small systems spanning at most millions of atoms. This means that any calculated properties will be statistically affected by the small size of such a system. For example, if walls are employed, any macroscopic observations will be affected by the atoms interactions with the walls. To solve this problem, periodic boundary conditions are used to approximate an infinite system in every direction. For example, any atom leaving the box on the right side reenters the system on the left side. The "systems" must therefore have the right shape which can be repeated infinitely without leaving any gaps between the system boundaries. The simplest box shape is triclinic which has the disadvantage, however, that when simulating a spherical object, it creates additional space in the corners of the systems. These spaces are most often filled with water which decreases the performance drastically. For this reason a dodecahedron box shape can be used which I do in Chapter 6 and 7.

2.3 Water & Protein Models

The available water models vary widely having one to six "sites" to help approximate the complex nature of water. These sites can either correspond to the atom nuclei directly, or behave like virtual particles to represent the charge distribution within the water molecule. The number of sites is a trade-off between the computational cost and the quality of reproducing the physical properties [129]. One of the more widely used models is the 3-site model called TIP3P [130] and it has been adapted in the CHARMM forcefield family [131]. However, the CHARMM

version of TIP3P differs in that it uses the Lennard-Jones parameters on the hydrogen atoms [116].

In this thesis I used the CHARMM family of forcefields. It is extended with one more potential term called CMAP. This new potential term is used by the CHARMM family in order for the results to be more consistent with other empirical forcefields. Particularly, it corrects the dihedral potential energy term of the protein backbones. The CHARMM36 forcefield builds upon the CHARMM22/CMAP which suffered from problems with, among other things, overestimating the helical content in proteins [131].

CHARMM36 forms the basis for the CHARMM36m forcefield [120]. The new refinement improves the ability of the forcefield to model intrinsically disordered proteins (IDPs). Intrinsically disordered proteins, rather than having a well defined structure, have site(s) which can assume multiple different conformations, or that do not have any well defined set of conformations, while still fulfilling a biological function. In order to model IDPs, the small angle X-ray scattering (SAXS) was used to estimate the dimensions of peptides. In addition to SAXS, fluorescence resonance energy transfer (FRET) is used to probe the dimensions of smaller peptides. It was found that CHARMM36 is underestimating the size of the protein ensembles. This problem, however, affects all forcefield families [132]. For example, the ensemble average radii of gyration (R_g) of retroviral integrase is approximately 24 Å rather than the modelled 13 Å [120]. Furthermore, CHARMM36 simulations of disordered arginine-serine peptide overestimated the population of the left-handed α-helix [131]. In C36m this was corrected which the group demonstrated using four IDPs, improving significantly the agreement with the available NMR and SAXS data.

Whereas for GB1 β-hairpin the quantity of β-hairpin has not changed across the two forcefields. However, CHARMM36m underestimates the stability of some β-hairpins, as NMR estimates of folded chignolin and CLN025 were found to be substantially larger. The consequence of this on fibronectin simulations is not easy to predict. Altogether, the new forcefield was validated using 15 peptides and 20 proteins with the total simulation time counting over 500 μs [120].

2.4 Analysis and Software

Molecular Dynamics simulations generate, among others, coordinates and velocities over time. The different time-points of a trajectory are referred to as frames. These have to be analysed, and many different analysis tools are available [133]. Frequently the very specific requirements and evolving technology necessitates rewriting the tools. Several analysis scripts that rely on the package MDAnalysis [134] have been written and the code is made available in Appendix A. Here, two tools used in the analysis are discussed, the Spatial Density Map and the clustering algorithm DBSCAN [135].

2.4.1 Spatial Density Maps

Spatial Density Map (SDM) is a three-dimensional heat-map of atomic positions relative to the area of interest. SDM can be used to visualise the hydration of a surface, or other elements in the system such as different functional groups in a protein. In order to create an SDM, one should define the frame-of-reference which can be used to superimpose the atoms in the system across time [136]. The frame-of-reference requires at least three suitable coordinates, which can be atoms A-B-C. These coordinates/atoms can be used to define the *xyz* of the system.

One way to construct the frame reference is to select one of the atoms as the origin, for example B. Then, the coordinates of the entire system in that frame should be translated such that B = (0, 0, 0) in the system. Then, the first vector $\vec{BA}$ can represent the x-axis. By taking the cross product of $\vec{BA}$ and $\vec{BC}$ we obtain a perpendicular vector to the plane ABC which will form the z-axis, and by taking the cross product of the z-axis again with the $\vec{BA}$ (or x-axis), we will obtain our y-axis. The y-axis should be relatively well aligned with the vector $\vec{BC}$. This approach is taken by [136].

Having the three defined points we can rotate the entire system such that the defined frame-of-reference are aligned with (1, 0, 0), (0, 1, 0), (0, 0, 1) which in most molecular dynamics simulations represent the three axis *xyz*. By constructing the axis for each frame and rotating the coordinates accordingly, we bring all the frames to the same frame-of-reference. After

the rotation, the initially selected A-B-C atoms should be superimposed across the frames. Similarly, any other atoms in the system will remain in the same relative position to the initially selected atoms. Therefore, averaging the density in the space over time can help understand where the atoms prefer to reside.

The construction of the frame-of-reference can be omitted, however. One can use the atoms directly. For example, having three atoms A-B-C, one can translate the system such that B is the origin, and then rotate the all the atoms in the system such that A and C points are as close as possible to their location in the first frame. However, in both cases it has to be possible to create the frame-of-reference.

The three points used as the frame-of-reference cannot be in a single line. In that case, it would be possible for the system to superimpose the atoms perfectly, but randomly, around the axis of the line. Furthermore, the three chosen atoms have to remain in the same relative positions with respect to each other over the time of the trajectory. Otherwise, the superimposition across the frames will be of poor quality.

2.4.2 Data Clustering

In order to analyse protein adsorption over time, the adsorption states should be grouped accordingly. For this, the data clustering can be used. Consider adsorption of a protein P to the surface S. In a simple case, the surface presents the same atoms (e.g. metallic surface), and the aim is to extract the different ways in which P can adsorbs to S. The adsorption of P on S can be captured by taking the distance from each atom in P to the closest atom in S. Therefore, for P with n atoms, one adsorption state is a n-long array of distances. As a consequence, a trajectory can be thought of as an array of n-long distances. We refer to any of the adsorption states as PS(t), which represents the distances of each atom in P to S at time t.

In order to compare the adsorption states i and j, we calculate their difference D_{ij}:

$$D_{ij} = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(PS_{ik} - PS_{jk})^2}$$

where PS_ik and PS_jk is a distance of the atom k to the surface at times i and j, and n is the number of atoms in P. The value of D_{ij} equal to 0 when the two adsorption states i and j are identical.

Once we have the metrics of distances between any two frames, we can use DBSCAN, which is one of the most popular clustering algorithms. DBSCAN is a non-parametric algorithm. It groups together closely packed adsorption states (many nearby neighbours), while also marking adsorption states that are not similar to any other adsorption states as noise (low-density regions).

DBSCAN takes one parameter, epsilon, which is the radius of a neighbourhood. This parameter is a starting point for it being able to classify points either as core points or density-reachable points, and outliers. In the case of the example used in the previous paragraph, 1 Åmeans that a protein can be differently positioned from the surface by 1 Åon average. Core points are points that around which more atoms can be found (within ϵ). A point is directly reachable if it is within ϵ of a core point. A point q is reachable from p if there is a path p_1, ..., p_n with p_1 = p and p_n = q, where each p_{i+1} is directly reachable from p_i. Note that this implies that all points on the path must be core points, with the possible exception of q. All points not reachable from any other point are outliers or noise points. Now if p is a core point, then it forms a cluster together with all points (core or non-core) that are reachable from it.

2.5 Computing power

The molecular dynamics software used in this thesis is the open source package GROMACS [137]. It is one of the fastest packages running on CPUs due to state of the art implementation of the SIMD instructions, which can be described as the "parallel units of a CPU". GROMACS implements a message-passing interface (MPI) which makes it possible to employ multiple computers or nodes simultaneously. I used this technology during the first half of my PhD after which newer GROMACS packages became available. The latest versions, including GROMACS 2018 and 2019, have shifted their focus to the graphics processing units (GPUs), offering very

good performance [138]. The rapid improvements in the software and in the GPU hardware have substantially decreased the financial cost of molecular dynamics simulations. This allowed me to increase the number of simulated systems in the last chapter of this thesis.

Without further ado, the first simulation results are presented in the following chapter.

Chapter 3

$FnIII_{9\text{-}10}$ Adsorption to EA and MA SAMs

Fibronectin (Fn) fibrillogenesis organised fn into fibrils which *in vivo* requires a cell-mediated cascade of events that relies on integrin binding and cytoskeletal reorganization, during which Fn transitions from a compact to extended conformation. Fn fibrillogenesis can also be induced with the right substrate, such as poly(ethyl acrylate) [PEA], on which Fn becomes extended, starting fibrillogenesis. Interestingly, the almost chemically identical polymer poly(methyl acrylate) [PMA], which has only one fewer methylene bridge ($-CH_2-$), does not lead to fibrillogenesis. To understand the cause of this difference in Fn behaviour on PEA and PMA, I modelled the two substrates using ethyl acrylate (EA) and methyl acrylate (MA) self-assembled monolayers (SAM). The domains $FnIII_{9\text{-}10}$ which are pivotal in the process of cell-mediated fibrillogenesis were atomically simulated for each SAM. A prompt and stable adsorption on EA SAMs and no adsorption on MA SAMs was observed. The analysis of the water hydration at the EA and MA SAMs shows that the extra methylene group in the EA functional group creates new degrees of movements, leading to a markedly less dense hydration. This less dense hydration affects fibronectin adsorption to the substrate which can affect the process of fibrillogenesis.

3.1 Methods

For the molecular dynamics (MD) simulations, the structures of the self-assembled monolayers (SAMs) molecules was drawn with the open source software Avogadro [139]. The chains are defined as $SH(CH_2)_nR$, where R is $-COOCH_3$ for EA and $-COOCH_2CH_3$ for the MA terminated SAMs (Figure 3.1a). Two chain lengths, $n = 10$ and 18, have been investigated for both functional groups in order to discuss sampling. I first report the findings from the $n = 10$ SAMs and later compare it to $n = 18$. The SAM chains were parameterised with CGenFF [140] for the all-atom forcefield CHARMM36 [131] with low penalties.

System	Tilt (°)
EA10	59.2
EA18	61.0
MA10	59.6
MA18	60.0

Table 3.1: The average tilt of the self-assembled monolayers (SAMs) after equilibrium. The last 10% of the datapoints were used to obtain the average tilt.

For each SAM, the GROMACS tool genconf [137] was used to assemble 500 chains into a grid with an area per chain of 21.5 $Å^2$. Each grid was then placed on a gold substrate with dimensions of 10.57 Å × 10.18 Å, which was modelled with the polarizable GolP-CHARMM forcefield [141] with the sulphurs 2 Å away from the substrate. Atoms in the gold slab were frozen in the all dimensions, except for the virtual atoms representing the charge. A steepest decent minimisation was carried out with a 0.01 nm initial step size and a target of 500 kJ mol^{-1} nm^{-1} maximum potential force before the equilibration simulations were started. The NVT ensemble, which denotes a constant number of particles, volume and temperature, was then used to equilibrate the system for 20 ns. SAMs equilibration for $n = 18$ follows a similar protocol. The increased chain length makes it more difficult for any of the SAM molecules to escape the monolayer. The SAM was constructed so that their sulphurs were 10 Å away from the substrate. Due to the longer length of the hydrocarbon chain, the initial distance of the sulphurs from the surface can be larger, as it is more difficult for any of the molecules to escape the SAM. Before the addition of water, the system was equilibrated for 1 ns. The system was then simulated in NVT for 5 ns in the presence of water. The created SAMs ended with a very

similar surface tilt, as shown in Table 3.1.

A 9 nm wide slab of water molecules modelled with the CHARMM36 TIP3P forcefield was added on top of the substrate. After the addition of water, the system energy was re-minimised and an NVT simulation of 1 ns was performed.

The $FnIII_{9\text{-}10}$ structure was extracted from the crystal structure (PDB 1FNF [1]) to include the residue identifiers 1327 - 1415 for the 9^{th} domain and 1416 - 1509 for the 10^{th} domain. The interactions of the peptide were modelled with the CHARMM36 protein forcefield. The two domains were placed in the centre of the simulation box in the xy plane and at a location in the z-dimension such that the closest atom in the protein was 20 Å from the SAM interface. The protein was oriented with its RGD motif facing away from the SAM surface. Water molecules within 2 Å of any atom in the protein were removed and the system was neutralized with one sodium ion (Na^+) due to the -1e charge of the 10^{th} domain. After removing the overlapping water molecules, each of the four systems contained between 144,000 and 152,000 atoms. Figure 3.1b shows an example of the initial configuration of the peptide and the SAMs (water is not shown so the protein can be seen).

The NVT ensemble was used because the parameterisation of the polarizable gold substrate, GolP-CHARMM forcefield, was carried out in the same ensemble. The simulation box is periodic in the x- and y-dimensions and carbon walls were added to stop molecules from crossing the box in the z-dimension. The Particle Mesh Ewald (PME) algorithm which correctly accounts for the slab geometry of these simulation boxes was used to account for the long-range electrostatics. The temperature (300 K) was controlled by a velocity rescaling algorithm [122] containing a stochastic term, with a timestep of 2 ps. The cutoff for the van der Waals and Coulombic interactions was set to 1.2 nm and the LINCS constraint was applied to hydrogen-containing bonds [128]. The GROMACS version 5.x MD package was used to simulate each of the four systems for 500 ns.

In order to investigate the hydration of EA and MA SAMs, additional simulations of the SAMs coated gold substrate were carried out for 50 ns following the same equilibration protocol.

Whereas I used classical molecular dynamics simulations to study the systems, it is possible to

improve the sampling with the use of enhanced sampling algorithms, such as Replica Exchange Molecular Dynamics as well as Simulated Annealing Techniques [91]. These methods dynamically change the temperature of the system in order to escape the local minima in the energy landscape.

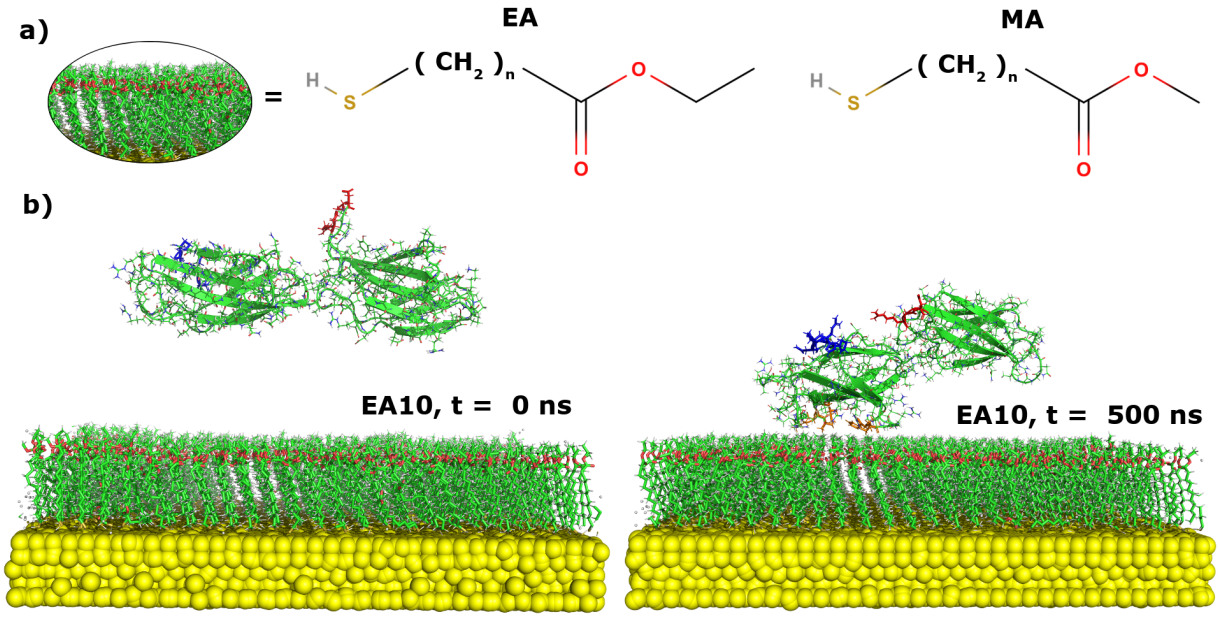


Figure 3.1: **a)** The structure of the four different molecules used to create the SAMs $n = 10$ or $n = 18$. In **b)**, a snapshot of the $n = 10$ EA SAM and the peptide at the starting time and at the end of the simulation ($t = 500$ ns).

Analysis The open source package MDAnalysis [142] was used to compute distances, coordination numbers, radial distribution functions (RDFs) and spatial density maps (SDMs). The electrostatic surface was computed with the PDB2PQR [143] online service followed by APBS [144] with default values and pH 7 for the two domains $\mathrm{FnIII_{9\text{-}10}}$. The SDMs were visualised with VMD [145, 146], the electrostatic maps were visualised with PyMOL [147], and other figures were created using Matplotlib [148].

The centre-of-mass (COM) distance was calculated for each domain with respect to the nearest heavy atom (non-hydrogen) in the SAMs. A distance below 20 Å generally means that the domain is in contact with the SAM interface. However, due to the ellipsoidal shape of the domains, a COM-substrate distance of less than 20 Å is not sufficient to determine whether the domain is in contact with the substrate. So in these cases any claims of contact have been verified with VMD.

The radial distribution function (RDF) describes the radial distribution of one group of atoms around another group of atoms with respect to the density of the system, $g(r) = \frac{\rho(r)}{\rho}$, where ρ is the global density of the object of interest and r the distance from the reference atom. To calculate the coordination number of oxygen atoms in the water molecules around a given atom, the average number of oxygens within a given distance of that atom was determined (e.g. the first neighbour distance as determined from the RDF) on the SAM over the course of the simulation.

Spatial density maps of the SAM hydration shells were created by superimposing the esters of EA and MA to first minimize the root mean square deviation. The rotational matrix transformation applied during the superimposition was also applied on the water oxygens found within a distance of 5.8 Å to the C20 atoms (see Figure 3.5). The distribution of the oxygen atoms in the water molecules was discretized within a three-dimensional grid and then normalised by the number of snapshots extracted before the generation of the density file. The isosurface shown in Figure 3.5b was created using the isovalue corresponding to the 95^{th} percentile of the density of water around the MA functional group.

The exposure of the RGD motif is described by the relative distance of the centre-of-mass of the motif and the centre-of-mass of the peptide, to the surface. Both domains were used for this calculation due to the position of the RGD motif on the loop between the two domains. A similar analysis was performed to determine the exposure of the PHSRN motif. However, in this case the centre-of-mass of the motif was compared to the centre-of-mass of the 9^{th} domain, as the PHSRN motif is found in that domain.

The potential interaction energy between the protein fragment and the surface was calculated with g_mmapbsa ([149]) using the molecular mechanics (MM) mode of the software to calculate the Coulomb and Van der Walls terms. Both of these terms were then visualised using a rolling mean and a rolling standard deviation with 1 ns window. The structural stability data (DSSP) was visualised with an in-house python script with matplotlib.

The python scripts used in the analysis and visualisation have been uploaded to github (`https://github.com/Lorenz-Lab-KCL`).

3.2 Results

Self-assembled monolayers (SAMs) were modelled using alkenothiols on gold and functionalised with ethyl acrylate, EA ($-C(=O)OCH_2CH_3$) and methyl acrylate, MA ($-C(=O)OCH_3$) to mimic the properties of the polymers PEA and PMA.

In the molecular dynamics simulations, domains $FnIII_{9\text{-}10}$ were simulated on the two EA and MA SAMs, with two replicas for each SAM, totalling four 500 ns-long simulations. The 9^{th} domain adsorbed to the EA SAMs but not to MA SAMs. However, the 10^{th} domain did not adsorb to either. The analysis of the adsorbing residues shows that the same residue regions drive the adsorption on EA and that the adsorption has, to a large extent, a hydrophobic nature.

3.2.1 Adsorption of $FnIII_{9\text{-}10}$ on EA and MA SAMs

In order to gain a more detailed understanding of how fibronectin interacts with the self-assembled monolayer interfaces, I have analysed the all-atom molecular dynamics simulations of the $FnIII_{9\text{-}10}$ domains with the EA and MA interfaces. First, I calculated the minimum distance between the centre-of-mass of each domain and the heavy atoms in the substrate (see Figure 3.2).

During the first half of the EA simulation the 9^{th} and 10^{th} domain make contacts with the substrate. In the second half of the simulation, the 9^{th} domain strengthens its adsorption while the 10^{th} domain diffuses away from the surface. In the MA system, however, there is no adsorption taking place. Despite several short contacts made, primarily by the 10^{th} domain, both domains diffuse away from the MA surface.

The adsorption of the 9^{th} domain is divided into two periods: ea10pI and ea10pII. The period ea10pI spans the time $t = 10 - 204$ ns during which the 9^{th} domain is in contact with the substrate but the fluctuation of the distance indicates lack of stability in the adsorption. At the end of the first period the domain loses contact with the surface for approximately 30 ns. After that the domain stably adsorbs to the surface, and remains so over the entirety of period

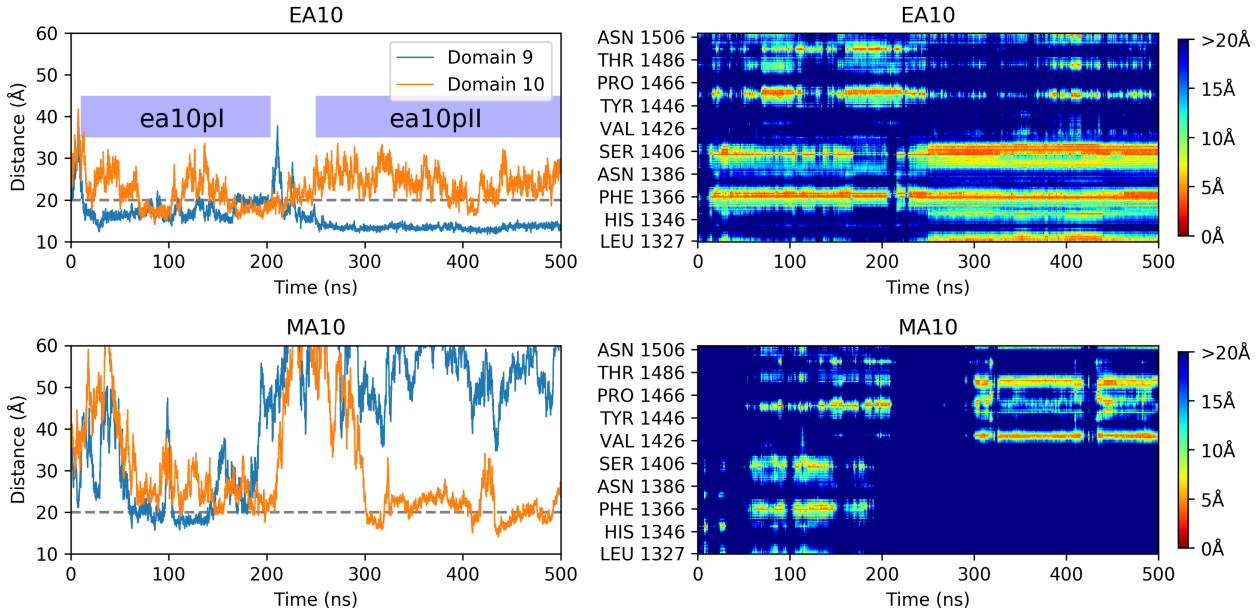


Figure 3.2: **Left)** The distances from the centre-of-mass of the 9$^{\text{th}}$ and 10$^{\text{th}}$ domains to the nearest heavy atoms in the SAMs, over time. At any distance over 20 Å (grey dashed line) the domain is unlikely to be in contact with the substrate. **Right)** The minimum distances between any heavy atom in each residue and any heavy atom in the SAMs over time.

ea10pII ($t = 250 - 500$ ns). In contrast, in the MA system, the 9$^{\text{th}}$ domain never adsorbs to the SAMs, making only two brief contacts with the MA10 substrate at $t = 60 - 93$ ns and $t = 104 - 148$ ns.

The 10$^{\text{th}}$ domain does not adsorb to the substrate in the EA10 system. However, it makes two contacts during the first half of the simulation at $t = 70 - 105$ ns and $160 - 220$ ns (yellow patches in Figure 3.2). During the first contact of the 10$^{\text{th}}$ domain, the 9$^{\text{th}}$ domain is weakly adsorbed but towards the end of the second contact, the 9$^{\text{th}}$ domain improves its adsorption. Once the 9$^{\text{th}}$ domain is adsorbed stably, the 10$^{\text{th}}$ domain predominantly stays away from the substrate. This suggests that the adsorption of the 9$^{\text{th}}$ domain adversely affects the adsorption of the 10$^{\text{th}}$ domain.

Key residues in the adsorption of $FnIII_{9\text{-}10}$ to EA SAMs

In order to understand the interactions with the surface I calculated the minimum distance between the heavy atoms in each residue of the Fn domains and the heavy atoms in the SAM molecules. In Table 3.2, for each identified simulation period, I list the key residues in the

SAM	Interval (ns)	Residues
	Domain 9	
EA10	ea10pI: 10 - 204	Phe1366, Ser1367, Arg1369
	ea10pII: 250 - 500	Leu1327, His1365, Phe1366, Glu1404 Ser1406, Pro1407, Leu1408, Ile1410
MA10	60 - 93	Phe1366
MA10	104 - 148	Phe1366
	Domain 10	
EA10	70 - 105	Gly1456, Asn1457, Ser1458, Pro1459
EA10	160 - 220	Asn1457, Ser1458, Pro1459
MA10	326 - 416	Pro1430, Thr1431, Pro1479
MA10	434 - 497	Pro1430, Thr1431

Table 3.2: Residues less than 6 Å away from the surface for at least 80% of the specified intervals. The blue shaded entry denotes the most stable adsorption period.

adsorption process extracted from the residue-surface distance maps.

There are three overall trends in the EA system (Figure 3.2). First, the domains interact with the interface via the same regions: the residues clustered around Phe1366 and Ser1406 in the 9^{th} domain, and those found around Gly1456 and Ser1496 in the 10^{th} domain. Secondly, more residues in the 9^{th} domain interact with the surface than from within the 10^{th} domain. Thirdly, there is a clear relationship between the increase in the number of interacting residues in the 9^{th} domain and the decrease in the 10^{th} domain, which suggests that interdomain interactions affect the adsorption of the 10^{th} domain.

The region surrounding Phe1366 has several residues which adsorb strongly to the surface. The fact that this strongly interacting residue is a phenylalanine suggests that hydrophobicity plays a role in the adsorption. During the stable adsorption period ea10pII, several residues of different moieties are found interacting with the surface. However, more than half of these are hydrophobic.

The 10^{th} domain contacts both EA substrates sporadically. Some residues in the 10^{th} domain, despite not qualifying as important residues by our criteria, show a propensity to stay close to the surface. Particularly Asn1457 which is even found close to the surface during the second half of the simulation. Other residues which tend to come into contact with the surface include Thr1454, Gly1455, Gly1456, Asp1482 and Asp1495. These recurrent contacts suggest that the 10^{th} domain does attempt to adsorb to the surface. While the residue Leu1327 at the

N-terminal of the peptide, adsorbs, this is not of biological significance because this region of the peptide would be inaccessible due to a short interdomain linker.

Interactions of the peptide backbone & sidechains with the surface

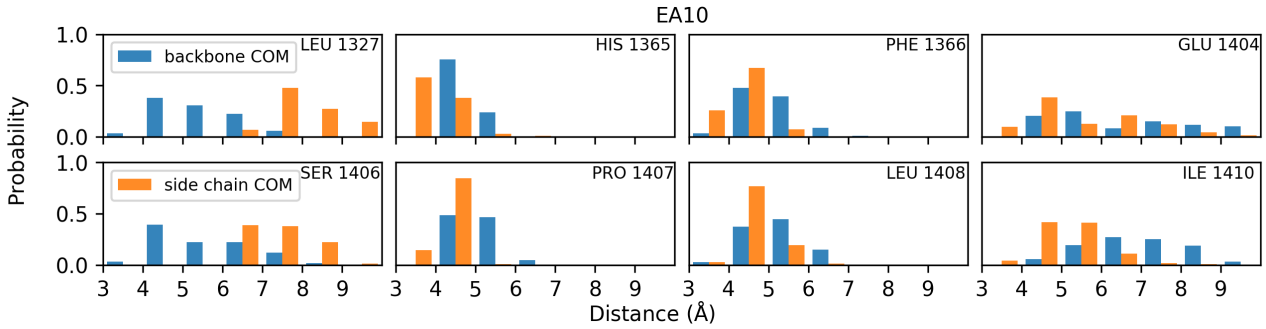


Figure 3.3: Histogram of the distances between the backbones and side chains of the selected residues and the functional groups on EA10 substrate. This analysis was applied only to the residues important in the adsorption of the 9th domain on EA10 (t = 250 - 500 ns), which have been previously listed in Table 3.2

In order to understand how the residues drive the adsorption of $FnIII_{9\text{-}10}$ to the EA10 SAMs, I check whether the residues interact via their side chains or their backbones by calculating the minimum distances between their centres-of-mass and the EA SAMs. To focus on the residues that contribute the most to the adsorption I consider only the 9th domain during the most stable adsorption period ea10pII. For each selected residue (Table 3.2), a histogram of distances is shown in Figure 3.3.

During the ea10pII stage five residues are close to the surface: His1365, Phe1366, Pro1407, Leu1408 and to a lesser extent Ile1410. Except for His1365, they are all hydrophobic. Phe1366 is found predominantly close to the substrate, with its side chain always found closer than its backbone. Each of the five residues prefers to interact via its side chain. Ser1406 has its backbone occasionally close to the substrate, but together with Glu1404, these residues show little specificity as their backbones and side chains are spread across the different distances. This trend suggests that it is the hydrophobic interactions that drive the adsorption. However, the positively charged His1365 has its backbone and side chain very close to the substrate at all times. I analyse the potential interaction energy between the protein and the SAMs to understand this better.

3.2.2 Energetics of adsorption of FnIII$_{9\text{-}10}$ to EA SAMs

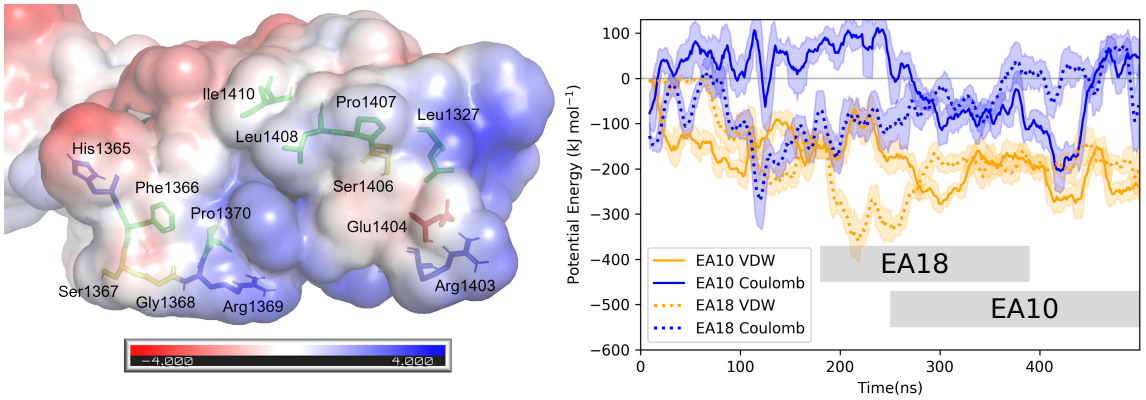


Figure 3.4: **Left)** The electrostatic potential of the 9th domain with the bound residues with the scale units in kT/e. **Right)** The van der Waals and Coulomb potential energy of the protein interaction with the SAMs. Both replicas are presented in this plot. The two grey blocks represent the most stably adsorbed periods ea10pII and ea18pII.

In order to better understand the effect of hydrophobicity in the adsorption process, I computed an electrostatic potential of the peptide (Figure 3.4, left) and superimposed the important residues on the structure.

The residues which I have identified as playing a key role in the adsorption mechanism are mostly found in the hydrophobic region between the negative (red) and positive (blue) regions. Rotation of the domain shows that behind His1365 and Ile1410 there is a more negatively charged patch, whereas on the opposite site, behind Arg1369 and Arg1403 there is a more positively charged region (not shown). Thus, despite ample opportunity for interaction with the polar patches, it is the hydrophobic site that was involved in the adsorption.

One location has more than one hydrophobic residue in the same area. This region includes Leu1327, Pro1407, Leu1408 and Ile1410 and it adsorbs strongly during ea10pII (Figure 3.4, left). Closely to the left of this residue area, the hydrophobic residue Phe1366 is found. Previous analysis of this residue showed that, on average, its side chain is particularly close to the surface. The phenylalanine is not alone as it is accompanied by Pro1370. However, this residue was not found to be as close to the surface as often.

I calculated the non-bonded potential energy terms consisting of the van der Waals (vdW) and Coulomb terms for the protein-surface interactions (Figure 3.4, right). The results are

presented for both replicas EA10 and EA18. The contribution of the vdW interactions to the total potential energy is larger than that of the Coulombic interactions across the two EA systems.

In the EA10 replica, the Coulombic interactions are unfavourable to adsorption during the first half of the simulation. This changes, but only temporarily. The vdW interactions in turn continue improving steadily until the middle of the simulation. After that they are stable and fluctuate around -200 kJ mol^{-1}. The situation is similar to the replica EA18, in which there appears to be an inverse relationship between the vdW and Coulombic interactions. During the first half of the simulation, the improvements in vdW go hand in hand with a decrease in Coulombic interactions.

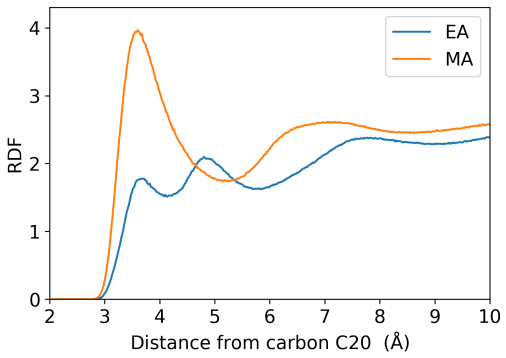
The Coulombic interactions are favourable to the adsorption in several places. For example, in EA10 they reach almost -200 kJ mol^{-1} at one point. However, they can also be unfavourable, as they are in both of the replicas towards the end of the simulations. This is in large contrast to the vdW potential energy which from the middle of both replicas steadily fluctuates around the value -200 kJ mol^{-1}.

It is therefore the vdW interactions between the residues and the EA substrates driving the adsorption which provides further evidence that the hydrophobic interactions play an important role in the adsorption of the 9^{th} domain to the EA SAMs.

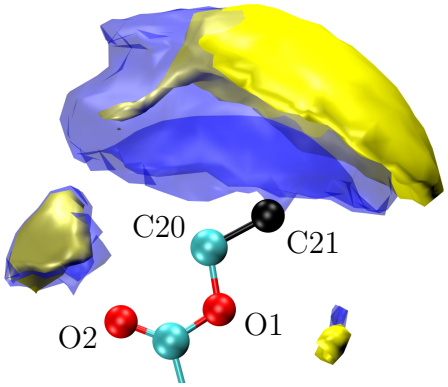
3.2.3 Hydration of the EA and MA SAMs

The loss of a single methylene bridge from the EA leads to a very different adsorption profile. Analysis of substrate hydration was performed for both MA and EA by calculating the radial distribution function (RDF) of the water oxygens around the SAMs double bonded oxygen (O2), single bonded oxygen (O1) and carbon (C20).

I present the RDF of the oxygen atoms in water molecules around the C20 carbon in the EA and MA functional groups in Figure 3.5a, where the difference in the hydration of these two terminal groups is apparent. The EA RDF has two distinctive hydration shells, with minima



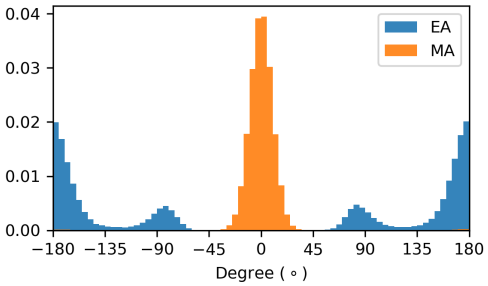
(a)



(b)

Atom	Distance (Å)	MA n_{Ow}	EA n_{Ow}
O2 - Ow	3.6	0.8	0.6
O1 - Ow	3.2	0.2	0.2
C20 - Ow	4.15	3.2	1.3
C20 - Ow	5.8	8.8	6.1

(c)



(d)

Figure 3.5: Hydration analysis: **(a)** RDF of the oxygen atoms in the water molecules from the C20 carbon in the SAMs, **(b)** Spatial Density Map of the oxygen atoms in the water molecules around the C20 atom for the MA (transparent blue) and EA (yellow) functional groups. The black atom represents a carbon atom present only in the EA functional group. Table **(c)** contains the nearest neighbour distance and coordination numbers for water molecules around various atoms within the EA and MA functional groups. The first two cut-off distances are based on the RDF of the water around the oxygen atoms (not shown), whereas the latter two are based on the RDF shown in a). The cut-off value 4.15 Å is the first minima of the EA, whereas 5.8 Å is chosen to describe both, the first two minima in EA and the first minima of the MA. Distribution **(d)** of the dihedral angle for EA (O2-C19-O1-C20) and MA (C19-O1-C20-C21) showing the rotation of the methyl group due to the extra methylene bridge.

at 4.15 Å and 5.8 Å which represent the hydration shell of the carbonyl O2 and that of the ethyl group C20-C21, respectively. In contrast, the first MA hydration shell is much denser, as denoted by the larger magnitude of the first peak of the MA RDF, whose width nearly spans the two EA's RDF peaks.

To compare the hydration of the EA C20-C21 and MA C20, I used a spatial density map to visualise their most dense hydration regions in Figure 3.5b. The part of the hydration shell that shows the largest density disparity between the two functional groups, as previously described, is on top of the C20 and C21 atoms. The dense part of the MA hydration shell (transparent blue) is much wider and covers the entire functional group. In contrast, the EA's hydration

shell (yellow) is smaller in size, and has a gap between the carbonyl oxygen O2 and the C21 carbon. This gap represents a region of lower density, increasing the opportunity to interact with the surface. Therefore, the MA functional group not only has a more dense hydration layer, but unsurprisingly, the visualised regions with the same density cut-off reveal that MA hydration shell is larger in volume.

To quantify the difference in hydration between the two functional groups I determine the coordination numbers (Figure 3.5c) by calculating the average number of water molecules found within a given distance. The most polar atom in EA, carbonyl O2, has a very similar coordination number to MA at the distance of 3.6 Å. The largest disparity is around the atom C20 where MA, at a radius of 4.15 Å, has on average 1.9 more water molecules than EA.

This difference between the EA and MA substrates is compounded by the extra methylene bridge in EA which leaves less space for water. However, even at the longer radius of 5.8 Å, I observe further increase in the difference in the hydration, despite this region not being affected by the extra atom. The difference in hydration at 4.15 Å is 1.9 water molecules, whereas at 5.8 Å the difference increases to 2.7 water molecules. Thus the presence or lack of this sole extra methylene bridge is not sufficient to explain the difference in hydration between the two substrates.

Looking at the different conformations assumed by the functional groups, due to the additional methylene bridge in the EA functional group, its methyl group rotates (3.5d). This freedom to rotate stops the ability of water molecules to form denser hydration shells. In contrast, the MA functional group stays in the same plane of the ester group in both EA and MA functional groups, as shown with the single peak of the MA. The additional degrees of freedom of the EA methyl group results in a significantly different hydration, explaining how despite having only subtly different chemistry, the adsorption profile of the $FnIII_{9\text{-}10}$ domains is markedly different on the two substrates.

The RDF, coordination number and spatial density maps show how much the hydration of EA and MA SAMs differ. Furthermore, it is shown how significantly a very small modification (an extra methylene bridge) in the number of degrees of freedom can affect complex systems

such as SAMs. Therefore, this substantial difference in hydration is likely to be an important component affecting the adsorption of the $FnIII_{9\text{-}10}$ domains.

3.2.4 Exposure of RGD and PHSRN motifs

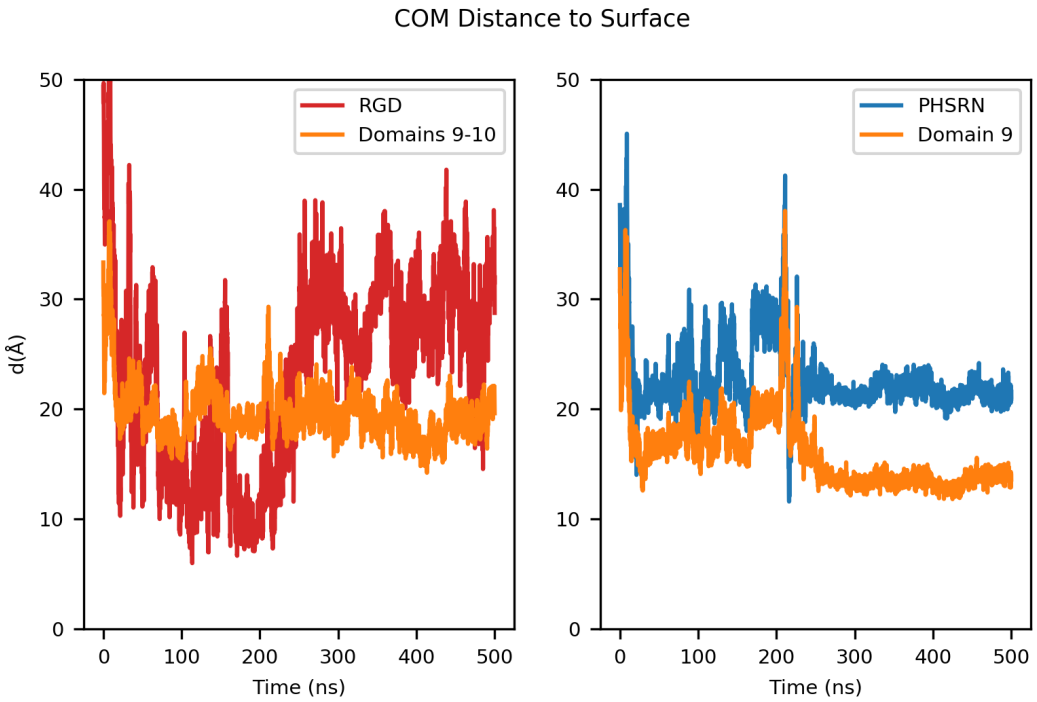


Figure 3.6: RGD and PHSRN motif exposure in the EA10 system. The shortest distances to the SAM (heavy atoms) from the motif centre-of-mass and the protein fragment centre-of-mass. The RGD loop, due to its location between the two domains, was compared to the centre-of-mass of the whole protein fragment (domains 9 and 10). The PHSRN was compared to the centre-of-mass of the 9th domain. When the protein is closer to the surface than the motif, the motif is likely exposed to potential interactions.

The RGD and PHSRN motifs can be either buried in the surface or displayed for potential integrin binding, which determines the cellular response. To test whether the motifs are displayed, I measure whether the centre-of-mass of each motif is closer to the surface than the centre-of-mass of the peptide (Figure 3.6). The centre-of-mass of each motif is generally farther away from the surface than the centre-of-mass of the peptide after the peptide has adsorbed to the EA surface (Figure 3.6a). Therefore, the two motifs are exposed for binding with integrin. This is particularly clear for PHSRN which is consistently on display due to the 9th domain being well adsorbed. RGD is similarly orientated away from the surface, although larger fluctuations are observed which I attribute to the flexibility of the loop F/G (See Figure 1.1).

3.3 Replicas EA18 & MA18

In order to ensure that the previously discussed series of events was reproducible, another set of simulations is used to replicate the findings. The new systems mimic the EA/MA SAMs. However, the chain length, instead of n = 10, is now n = 18. Despite this difference, the equilibration of the SAMs finishes with a similar tilt (Table 3.1). Therefore it is justified to refer to the length n = 18 as another replica because the basic interface presented by the SAMs to the peptide/water will be similar. Here, I present the results of the EA18 and MA18 simulations by comparing them to the results of the EA10 and MA10 systems.

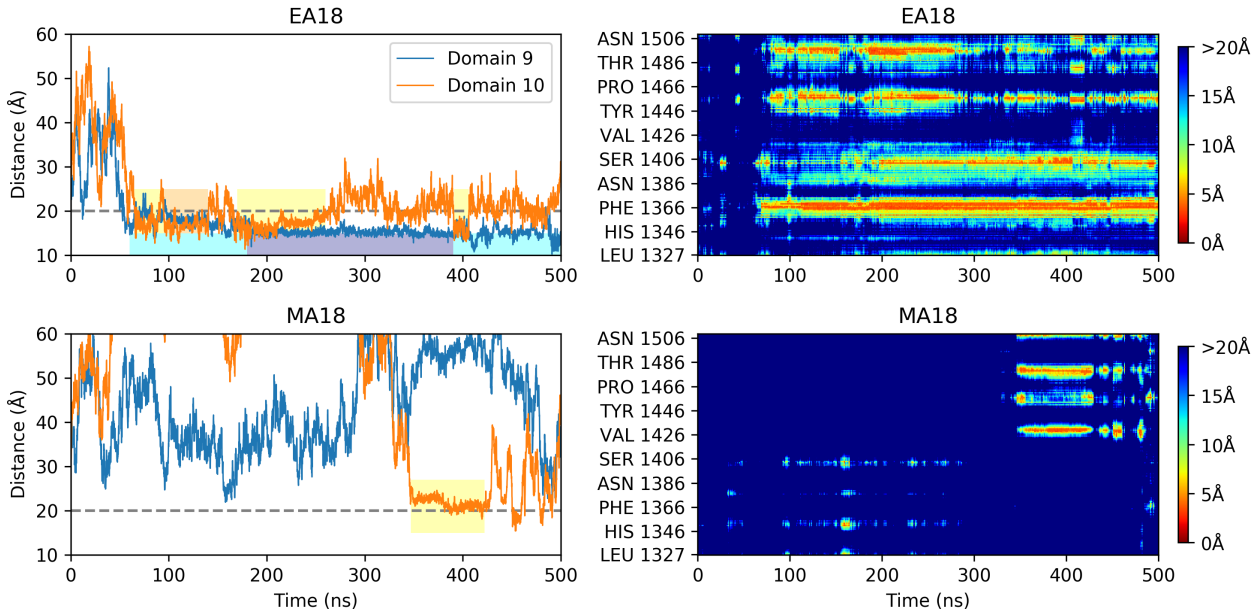


Figure 3.7: The graphs on the **left** visualise the distance from the centres-of-mass of the 9th and 10th domains over time to the nearest heavy atom in the SAMs. At distances over 20 Å (grey line) the domain is unlikely to be in contact with the substrate. The adsorption stages of the 9th domain are represented with blue-shaded patches. For the 10th domain yellow-shaded patches were used. The two graphs on the right represent the minimum distances between the residues and the interface of EA18 SAM over time. Heavy atoms were used for the distance calculations.

Adsorption The adsorption in the replicas EA18 and MA18 is consistent with the previous results. On EA18, the 9th domain similarly drives the adsorption, which in turn leads to the desorption of the 10th domain. It has been noted, however, that the 10th domain continues making contacts with the surface. The MA18 system follows the same behaviour as well, with

contacts made by the 10^{th} domain which are followed by the domain diffusing away from the substrate (Figure 3.7).

I again use the 9^{th} domain to divide the adsorption behaviour into separate periods which I name ea18pI, ea18pII and ea18pIII (blue patches in Figure 3.7). During ea18pI (t = 60 - 180 ns) the domain stays in contact but the distance to the surface fluctuates noticeably. The fluctuation disappears during ea18pII (t = 180 - 390 ns) which means that the adsorption of the 9^{th} domain becomes stable. After 390 ns, during ea18pIII, the domain-surface distance of the 9^{th} domain starts to oscillate about 15 Å away from the surface. The destabilisation of the binding does not lead to loss of contact with the surface, however. On the contrary, during this period, there are multiple occasions when the distance to the surface is smaller than observed in the second period ea18pII.

The 10^{th} domain in the EA18 system also follows the same pattern seen in EA10. Multiple contacts take place during the first half of the simulation at t = 65 - 90 ns, 95 - 140 ns and 170 - 260 ns as shown by the yellow patches in Figure 3.7. However, when the 9^{th} domain adsorbs stably and transitions to the ea18pII (t = 180 ns), the 10^{th} domain starts moving away from the substrate. After this loss of contact with the surface the 10^{th} domain makes one more contact at t = 391 - 406 ns which coincides with the disturbances in the adsorption of the 9^{th} domain.

Similarly to MA10, but in a stark contrast to both EA replicas, during the MA18 simulation, neither of the domains adsorbs stably. Both domains continue diffusing away from the surface. The 10^{th} domain makes a longer-lasting contact with the surface during the time $t = 347 - 422$ ns (yellow patch in Figure 3.7). However, this contact did not lead to adsorption.

Key Residues In Table 3.3 residues found close to the substrate during the identified intervals and contacts are listed.

First I compare the residues involved in adsorption to the EA replicas. The adsorption across the two SAMs relies on similar regions with a few overlaps. One overlap is the His1365 and Phe1366 which are common across the two strongly-adsorbed periods ea10pII and ea18pII. Furthermore, Phe1366 is the only residue that adsorbs strongly to the substrate during every

SAM	Interval (ns)	Residues
Domain 9		
EA18	ea18pI: 60 - 180	Phe1366, Ser1367, Arg1369
	ea18pII: 180 - 390	His1365, Phe1366, Ser1367, Gly1368, Arg1369, Pro1370, Arg1403
	ea18pIII: 390 - 500	Phe1366, Ser1367, Gly1368, Arg1369, Pro1370, Arg1403
Domain 10		
EA18	95 - 140	Asn1457, Pro1497
EA18	170 - 260	Asn1457, Ser1458, Asp1495, Ser1496, Pro1497, Ala1498
MA18	347 - 422	Pro1430, Pro1479

Table 3.3: Residues that are less than 6 Å away from the surface for at least 80% of the found contacts.

period across the two EA replicas.

There are other residues consistently close to the surface in the EA18 replica. These include Ser1367 and Arg1369 in addition to Phe1366. Furthermore, the latter two adsorption periods include Gly1368 and Pro1370. Interestingly, in the last period ea18pIII, the residue His1365 is not found near the surface. This suggests that its previous adsorption was due to being a neighbour of a strongly adsorbing residue Phe1366.

During the MA10 simulation two contacts were made by the 10^{th} domain at t > 300 ns (Table 3.2). In these contacts two residue regions were involved, both of which are also found engaged in MA18 (Table 3.3). The two sites are represented by the residues Pro1430 and Pro1479 in the MA18 replica. These two residues (and sites) are placed on the N-terminal end of the 10^{th} domain, just next to each other. In the MA10 system, this region was responsible for only one contact spanning $t = 347 - 422$ ns, which also is followed by desorption of the domain.

In addition to the two discussed residues, there are several residues in the 10^{th} domain that make sporadic contacts. These include Thr1454, Gly1455, Gly1456, Asp1482 and Asp1495. The recurrent nature of these contacts suggests that the 10^{th} domain continually attempts but fails to adsorb to the surface.

Backbone/Sidechain Preference During the period ea18pII the residues closest to the surface are Phe1366, Gly1368, Arg1369 and Pro1370. Two of these, Phe1366 and Pro1370 are hydrophobic. The short distance of 3 - 4 Å seen in Gly1368 is suspected to be due to its position between the other strongly-adsorbed residues. Positively charged side chains of Arg1369 and

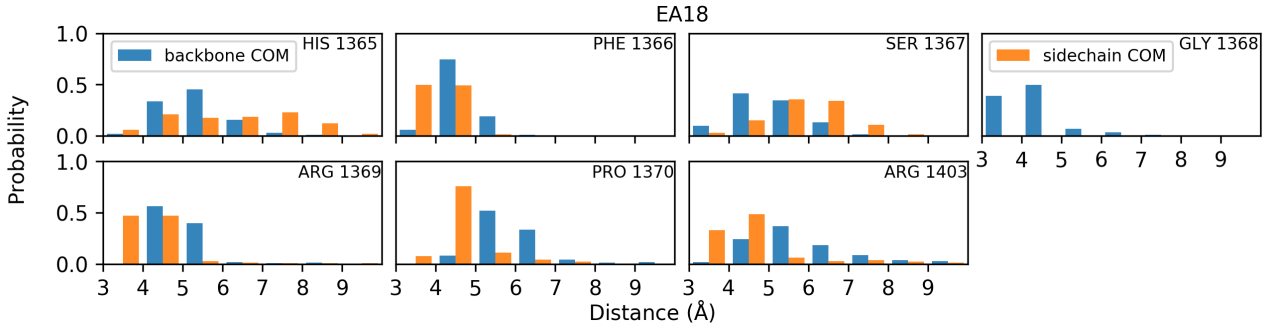


Figure 3.8: Histogram of the distances between the backbones and side chains of the selected residues and the EA18 surface. This analysis was applied only to the residues important in the adsorption of the 9th domain on EA18 (t = 180 - 390 ns), which have been previously listed in Table 3.3

Arg1403 are found relatively close to the substrate, but only Arg1369 is close to the surface throughout the whole period. The last residues, His1365 and Ser1367 rarely interact with the surface.

Across the EA replicas, besides the residue Phe1366, there are four other hydrophobic residues which adsorb well to the surface. The only charged residues that can be distinguished is Arg1369. The lack of specific interactions again suggests the importance of hydrophobicity in the adsorption.

Between the two EA periods I found only two common residues, His1365 and Phe1366. Phe1366 is found predominantly close to the substrate in both systems, with its side chain always found closer than its backbone. The second common residue, His1365, has its backbone and side chain very close to the surface in the EA10 replica. However, in the EA18 system, this residue mostly interacts with the surface via its backbone. This serves as additional evidence that His1365 is close to the substrate due to the surrounding residues, most likely Phe1366.

Exposure of RGD and PHSRN motifs The binding availability of RGD and PHSRN motifs in the EA18 replica was calculated as described before and is presented in Figure 3.9. The RGD is more often closer to the surface than the $FnIII_{9\text{-}10}$ domains. However, in the second half of the simulation, there is a lot of fluctuation in its distance to the surface. This is once again attributed to the flexibility of the loop F/G (See Figure 1.1).

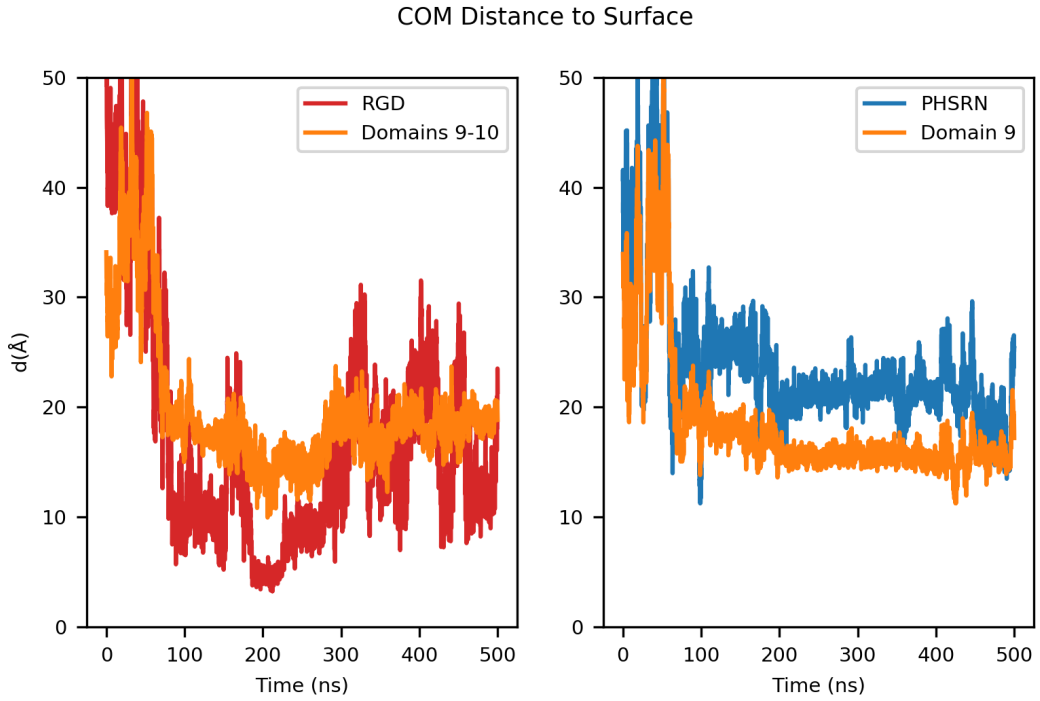


Figure 3.9: RGD and PHSRN motif exposure in the EA18 system. The shortest distance from the centre-of-mass of each motif and the corresponding protein domain(s) to the surface SAMs. The RGD loop, due to its location between the two domains, was compared to the centre-of-mass of the two domains $FnIII_{9\text{-}10}$. The PHSRN was compared to the centre-of-mass of the 9^{th} domain. When the protein is closer to the surface than the motif, it means the motif is more likely to be available for interactions.

In comparison to the EA18 replica, the exposure of the RGD motif is similar. The fluctuation shows that the motif is often presented on the domains. However, it is, on average, slightly less exposed for potential binding than in the EA10 replica.

The PHSRN motif is consistently exposed for binding in the EA18 replica. Towards the end of the simulations, however, it appears that the centre-of-mass of the motif and the centre-of-mass of the 9^{th} domain converge to the same distance. This is in contrast to the replica EA10 where the motif stays very well exposed during the second half of the simulation. Whereas further simulations are recommended to clarify this situation, the two motifs are on average well exposed for potential bindings.

The differences in motif exposure between the replicas are small. However, they show that despite the many similarities across the two replicas, the 9^{th} domain has not converged to the same adsorption state.

3.4 Stability

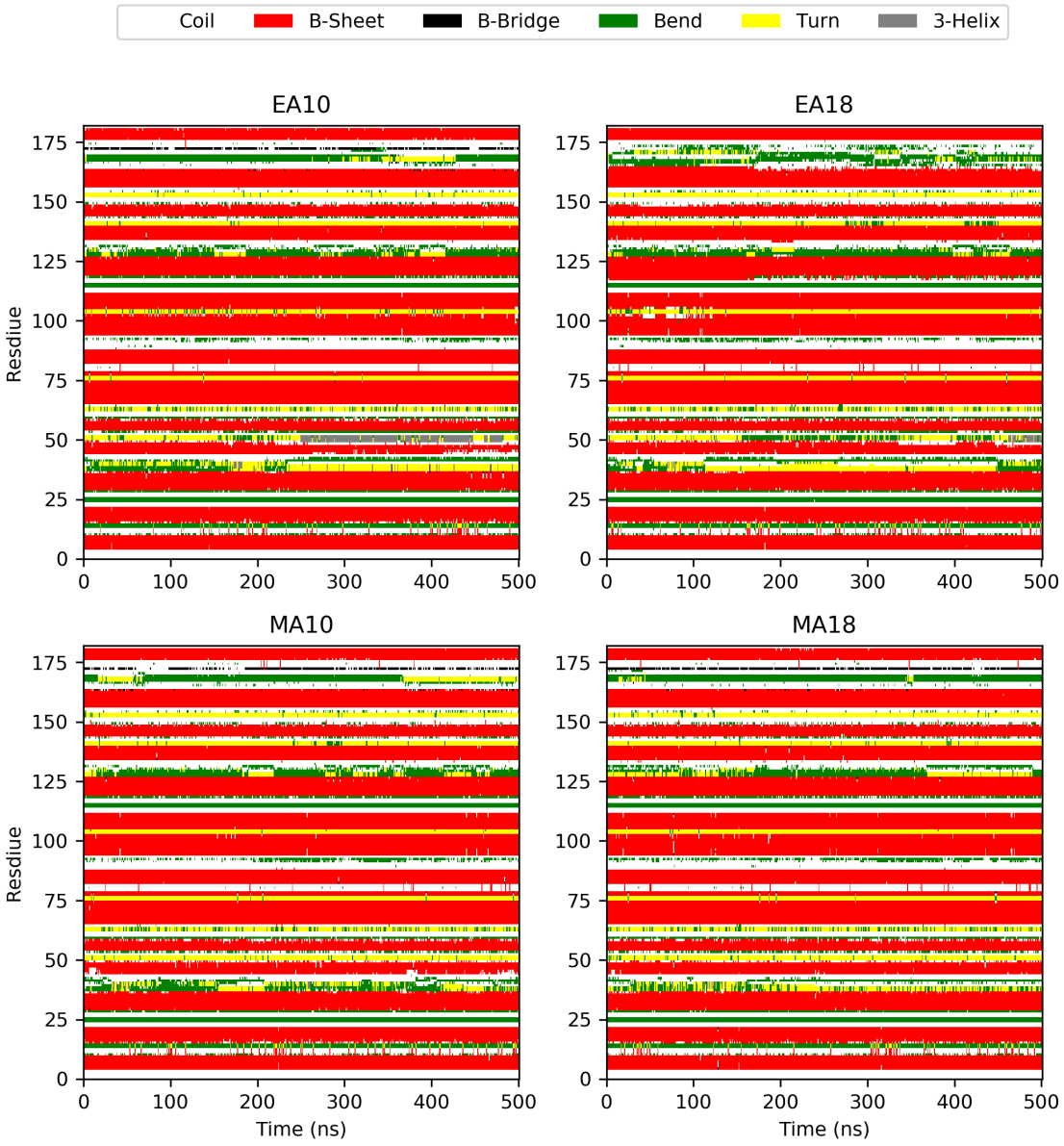


Figure 3.10: The secondary structure information computed with GROMACS DSSP and visualised using matplotlib. The secondary structures are highly conserved and very few changes are observed throughout the simulations.

The domains stability has been checked with DSSP software with the results presented in Figure 3.10. DSSP uses the atomic position information to classify each residue into one of the secondary structure categories such as β-sheet.

The red colour represents the residues classified as β-sheet. β-sheets are an important part of

the tertiary structure of the FnIII type. This conservation means that the domains are very stable across the four simulations and there is no indication of structural breakdown. Most of the observed changes are transitions from bend to turn and *vice versa* which is attributed to the loops and their relative freedom to move. They do not however affect the tertiary structure. Rather, they convey small changes to the way the β-sheets are connected.

3.5 Discussion

An additional methylene bridge in the polymer poly(ethyl acrylate) leads to fibronectin fibrillar network formation. In comparison, fibronectin on the polymer poly(methyl acrylate) appears to aggregate. This aggregated fibronectin, in the form of lumps, shows significantly less biological activity. In this work I modelled the adsorption of the two domains to EA-functionalised and MA- functionalised SAMs, which are used as simplified models of the PEA and PMA polymeric substrates, respectively.

This work was done in collaboration with the group of Manuel Salmerón-Sánchez who used atomic force microscopy to confirm that the SAMs models can replicate the described phenomena. Part of their results consisting of the atomic force microscopy results are shown in Figure 3.11 which will be available in our manuscript that has been accepted for publication in *Advanced Theory and Simulations*. Fibronectin formed aggregates on the polymer poly(methyl acrylate) as well as on the MA SAMs. On the other hand, fibronectin networks formed on the polymer poly(ethyl acrylate) and on its model of EA SAMs. The EA and MA SAMs reproduced the fibronectin behaviour and therefore can be used to simplify the polymeric substrates.

The simulations show that the adsorption of $FnIII_{9\text{-}10}$ takes place on the EA SAM, on which the fibrillogenesis takes place. However, no adsorption took place on the MA SAM (Figure 3.2) on which fibrillar networks are absent. The adsorption of the two domains on EA SAMs shows that both RGD and PHSRN motifs are exposed (Figure 3.6).

The simulations show that adsorption is mostly due to hydrophobicity, where the most contributing potential interaction energy term is the non-specific van der Waals. This is due to

a hydrophobic corridor in the surface of the protein (Figure 3.4) that drives the adsorption. Two simulations were carried out for EA and MA SAMs using different chain lengths (n = 10 and 18, Figure 3.1) making them similar enough to function as replicas, helping to account for sampling. Across the replicas, the adsorption, motif exposure as well as the interaction potential energy on the EA and MA SAMs followed the same patterns, showing that this is not a one-off event.

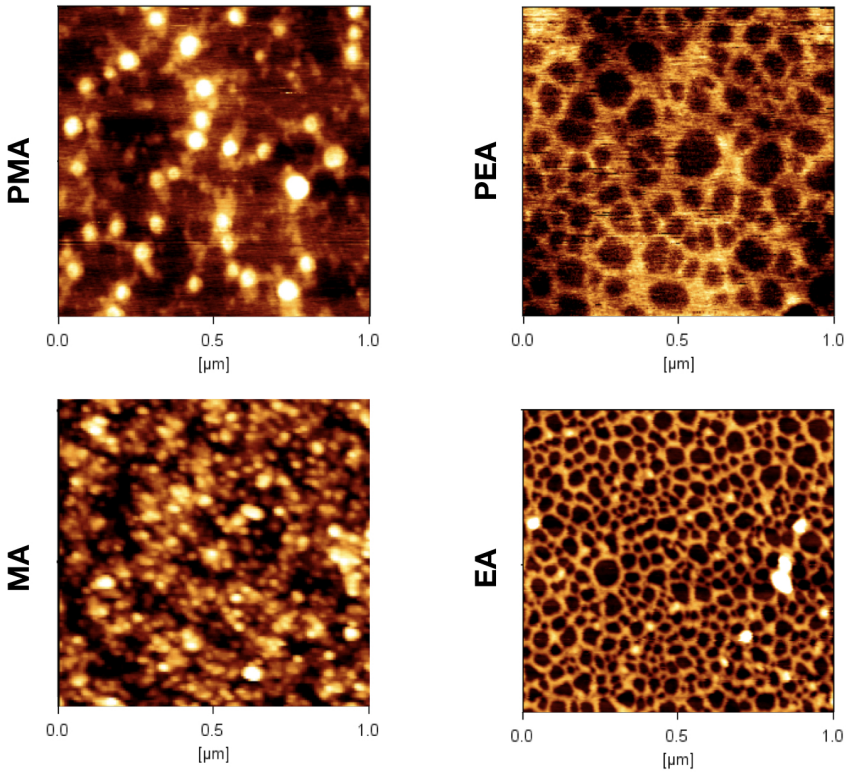


Figure 3.11: The Atomic Force Microscopy images of the fibronectin adsorbed on the different materials. A 20 $\mu g/ml$ fibronectin solution was used to perform the coatings. Height signal is represented. Work carried out by Virginia Llopis Hernández from Manuel Salmerón-Sánchez's group in Centre for the Cellular Microenvironment, University of Glasgow.

The simulations show that MA SAMs exhibit significantly denser local hydration shells than EA SAMs. I visualised the hydration to show that besides being denser, MA SAMs hydration shells are also larger in volume. I showed that the single extra methylene bridge introduces additional degrees of freedom in the EA functional group, allowing the methyl group to rotate which disrupts the hydration shell. In contrast, the MA functional group stays in a single plane

and allows for better ordering (Figure 3.5). I conclude that fibronectin is less likely to interact with the MA surface for this reason. The experimentally observed fibronectin aggregates or lumps might be explained with Fn-Fn interactions being stronger than the Fn-MA SAMs interactions.

The secondary structures in both domains remained very stable throughout the simulations (Figure 3.10). It is possible that changes in secondary structure are not necessary for the process of material-driven fibronectin fibrillogenesis. However, the simulation timescale is likely too small to draw this conclusion with any certainty.

It is important to acknowledge the sampling limitations in this work due to the computationally prohibitive nature of classical molecular dynamics. It is possible that extending the simulations further would lead to adsorption, nullifying the currently available results. For example, given enough sampling, the 10^{th} domain might be found to adsorb well to the surface after sufficient time, which could be a common result.

The simulations showed that the 9^{th} domain drives the adsorption of $FnIII_{9\text{-}10}$ fragment, leading to RGD exposure. This could offer another pathway through which the 9^{th} domain can affect RGD motif exposure - by modifying the availability of the motif through its own adsorption. However, the 10^{th} domain has showed clear signs of trying to adsorb to the surface. Furthermore, the adsorption strength of the 9^{th} domain correlates with the loss of adsorption in the 10^{th} domain. Assuming that the 10^{th} domain can adsorb to the surface, it is possible that with the right rearrangement of the two domains with respect to each other they would both be able to adsorb to the EA SAMs. For these reasons the adsorption of the 10^{th} domain in the absence of the 9^{th} domain is analysed in the next chapter.

Chapter 4

Adsorption of $FnIII_{10}$ on EA SAM

In Chapter 3 I presented how the $FnIII_{9\text{-}10}$ domains adsorb to the EA SAMs surface via the 9^{th} domain. Despite the 10^{th} domain making several contacts in the first halves of the simulations, it kept diffusing away from the surface. I further showed that there appears to be a relationship between the adsorption of the 9^{th} domain, and the inability of the 10^{th} domain to make contacts. Residue analysis further showed that the 9^{th} domain appeared to adversely affect the adsorption of the 10^{th} domain. In this chapter, the adsorption of the 10^{th} domain is investigated on the EA SAMs in the absence of the 9^{th} domain. Furthermore, I compare the exposure of the RGD motif to the availability of the motif in the previous chapter where the two domains were simulated together.

4.1 Simulations

The same protocol from Chapter 3 was used for the simulations of the 10^{th} domain. The software Avogadro [139] was used to construct the SAM chains defined as $SH(CH_2)_n - COOCH_3$. The EA SAMs mirror the previously defined systems with the chain lengths of n = 10 and 18. The SAMs were placed on a golden slab which is modelled with a polarizable GolP-CHARMM forcefield [141]. The SAM chains were parameterised with CGenFF [140] for the all-atom forcefield CHARMM36 [131], which was used for the protein.

The xy dimensions of the systems were reduced. After equilibration, each dimension was longer than 75 Å to ensure that the 10^{th} domain does not contact itself through the periodic box. This distance allows for plenty of space around the approximately 42 Å long domain. Each of the two final systems comprised around 60 k of atoms. The smaller size of the systems translated to better performance and made it possible to increase the sampling time to 1 μs for both replicas.

4.2 Adsorption

The two simulations of $FnIII_{10}$ are analysed using a similar protocol to the one used in Chapter 3. First, the distance from the domain to the surface is measured over time (Figure 4.1). The 10^{th} domain adsorbs to the surface within the first 100 ns across both replicas. After the adsorption takes place, the 10^{th} domain never leaves the surface.

In the EA10 replica, the adsorption of the domain quickly stabilises at around 12 Å away from the SAM (t = 110 ns) and does not change significantly during the remainder of the simulation. In the EA18 replica, the domain adsorbs in around 85 ns time. However, the distance from the domain stabilises only in the middle of the simulation (t = 530 ns), after a small event where the domain moves away from the substrate for a short period of time, which is estimated to last around 35 ns.

Therefore, the 10^{th} domain adsorbs stably by itself, as opposed to when it is in tandem with

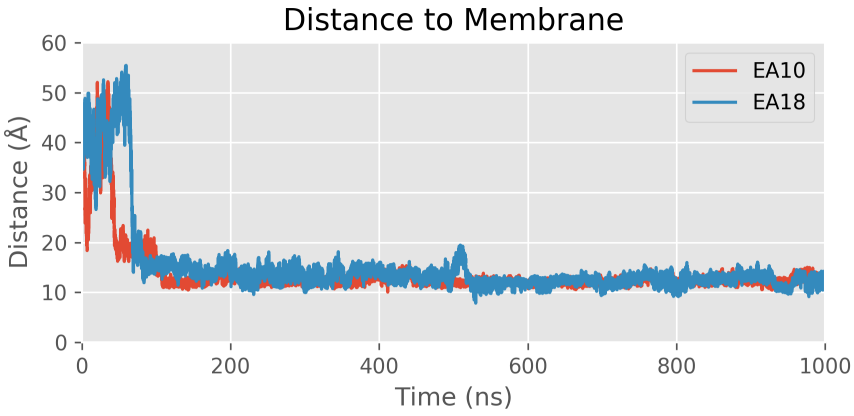


Figure 4.1: The adsorption of the $FnIII_{10}$ domain described by the nearest distance from the center-of-mass of the 10^{th} domain to the EA SAM over time. In both replicas prompt adsorption is observed, with the EA18 settling in the middle of the simulation.

the 9^{th} domain. It is possible that the residues important in the adsorption of the 10^{th} domain are made less accessible by the 9^{th} domain. To answer this question a closer look is taken at the residues that adsorbed, and whether these residues are blocked when the two domains are together. For the analysis I am going to use the intervals t = 110 ns - 1 μs for EA10, and 530 ns - 1 μs for EA18.

4.2.1 Residues

The nearest distances between each residue and the surface was calculated for the two replicas (Figure 4.2). A significant number of residues are involved in the contact with the surface in both replicas. Interestingly, different amino acids are close to the surface in each replica. In the EA10, there are four consistent regions that are strongly bound, and one that is weakly bound. The largest region spans residues 1460 - 1470, followed by regions centred around the residues Asp1495, Ser1475, Arg1445 and Pro1430. On the other hand, in the EA18 system, the adsorption of the 10^{th} domain is largely due to the residues Val1425 - Ile1435, and several dynamic and evolving contacts between Val1465 - Ser1475.

Whereas the same residues are used in the adsorption in the EA10 replica throughout the simulation, the picture in the EA18 replica is more complex. Initially, a single strongly bound region of residues stands out in the adsorption, spanning the residues Val1425 - Ile1435. Another

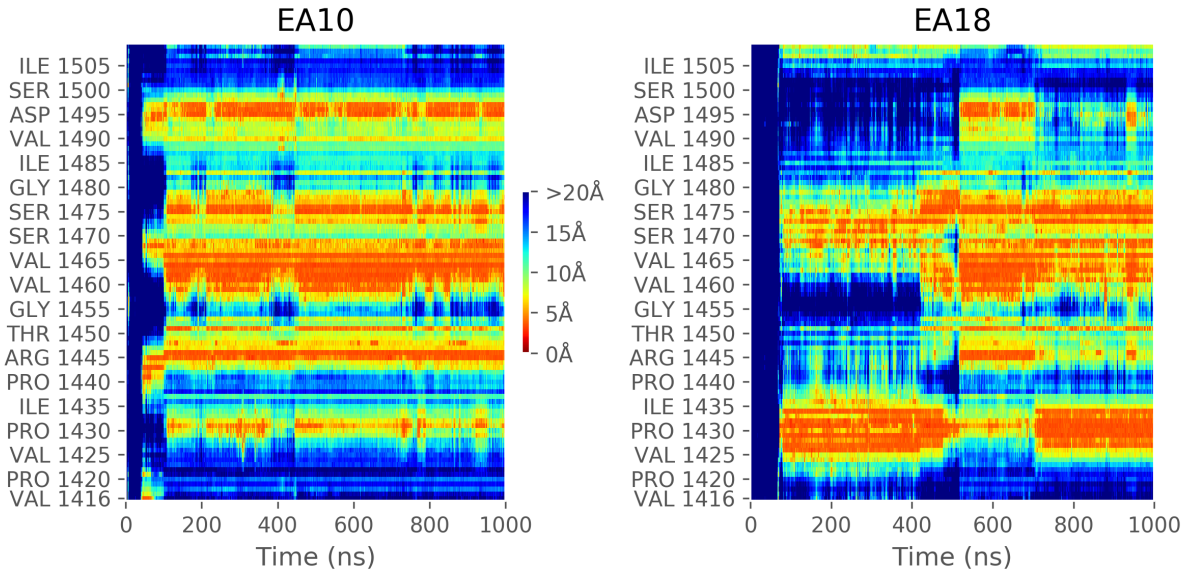


Figure 4.2: The minimum distance from each residue to the the surface over time. The adsorption in the EA10 replica does not change significantly, whereas in the EA18 replica, clear evolution of the adsorption to the surface is visible in the middle of the simulation.

region is centred around Ser1470 and it undergoes large changes during the adsorption, with a major breakthrough happening at 420 - 760 ns. This is the same rearrangement event found earlier in the COM-surface distance of the domain, in Figure 4.1. With the distance map presented here, the event can be discussed in more detail. First, the region Val1465 - Ser1475 starts to involve more residues and diverges into two separate regions after 400 ns time, expanding the scope of residues to almost Gly1455 - Gly1480. During this period, unexpectedly, two other regions become involved - these match the same regions found in EA10: the regions surrounding Asp1495 and Arg1445. Furthermore, at the same time, the large region surrounding Pro1430 looses many contacts, mirroring the behaviour of the region in EA10. In other words, the adsorption in the EA18 replica for almost 200 nanoseconds assumes the adsorption profile found in the EA10 replica. However, after that, the domain partly reverts to the previous configuration: the regions around Asp1495 and Arg1445 are mostly lost, the one around Pro1430 reverts to being the largest region. However, the Ser1470 does not return to its original profile: it is divided into two regions, which are similar to the regions in EA10 in the corresponding regions. Therefore it is concluded that partial convergence took place.

After the rearrangement event, after 720 ns, the two regions between Val1460 to Gly1480 in

the EA18 follow a similar path to EA10. Based on this the period of stable adsorption is updated. The coarser analysis in the previous section that relied on the COM_{Domain}-surface did not capture the transition in adsorption. Therefore, **for further analysis the time period 720 - 1000 ns is used** to extract the residues to which the adsorption converged.

The residues found in the last stable adsorption stages are extracted and listed in Table 4.1. The Table contains various types of residues including several positively and negatively charged, polar, and hydrophobic residues (Phe1463, Tyr1446), although the majority are charged or polar. In EA18, similarly, there is no clear pattern observed that would indicate whether the adsorption is of electrostatic of hydrophobic nature. It is worth noting that this is made more complex by the composition of the residues on the surface of the domain. The nature of tertiary structures stabilised by a hydrophobic core necessitate a number of polar or partially charged residues on the surface of the domain.

SAM	Interval (ns)	Residues
EA10	110 - 1000	Arg1445, Tyr1446, Gln1461, Glu1462, Phe1463, Thr1464, Val1465 Pro1466, Ser1475, Gly1476, Asp1495, Ser1496, Pro1497
EA18	720 - 1000	Ala1427, Ala1428, Thr1429, Pro1430, Thr1431, Ser1432, Leu1433 Leu1434, Pro1466, Lys1469, Thr1473, Ser1475, Gly1476

Table 4.1: Residues in the 10th domain that are less than 6 Å away from the surface for at least 80% of the final stably adsorbed periods.

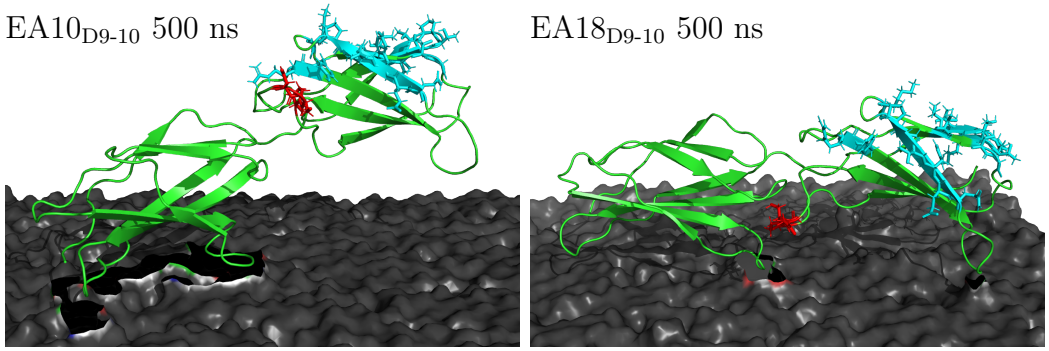


Figure 4.3: The last frames from the $FnIII_{9\text{-}10}$ in tandem simulations. Residues visualised as sticks are close to the surface during the adsorption of the sole 10th domain (4.1). The red colour highlights the Asp1495-Ser1496-Pro1497 which partly overlap with RGD motif (Asp1495). Others residues are coloured cyan.

Comparison to Tandem $FnIII_{9\text{-}10}$ I further compare the adsorption of the 10th domain to the short contacts made by the domain in tandem $FnIII_{9\text{-}10}$ in Chapter 3. Some residues in the

tandem 10th domain are very close to Glu1462 that featured on the list for the sole 10th domain. In the two $EA10_{9\text{-}10}$ contacts, there were four residues with the ID range 1456-59 (Table 4.1), which are spatially close. However, these exact residues were not found close the the surface in the adsorption of the sole 10th domain.

Let us investigate whether the replica $EA18_{9\text{-}10}$ behaves differently. In that tandem system, similarly to $EA10_{9\text{-}10}$, two residues close to the surface were Asn1457 and Ser1458, as well as residues 1495-98. Out of these three overlap with the residues that are close to the surface in the sole $EA10_{10}$ system: Asp1495-Ser1496-Pro1497. The involvement of these residue might explain why, in the $EA18_{9\text{-}10}$ system, as opposed to the 9th domain inhibiting the adsorption of the 10th domain, the 10th domain affects the adsorption of the 9th domain. In that simulation, towards the end, the two domains seem to compete for binding to the surface (Figure 3.7). This is likely related to the residue Asp1495, which is part of the RGD motif. The residues in the region adsorb well in the sole 10th simulations - and their contacts with the surface in the tandem $FnIII_{9\text{-}10}$ could cause the small perturbations in the $COM_{9,\ 10}$-surface distance, present towards the end of the simulation.

4.3 Potential Energy of Adsorption

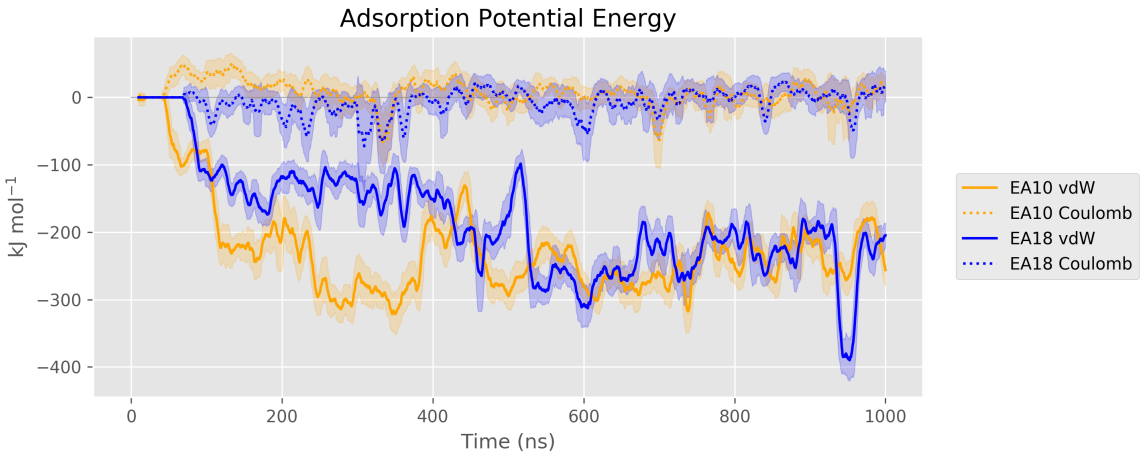


Figure 4.4: van der Walls and Coulomb electrostatic potential energy terms of the $FnIII_{10}$ domain interaction with the EA SAM. The rolling mean and standard deviation were used with the window interval of 1 ns.

In this section the nature of the interactions between the domain and the surface is investigated.

The electrostatic contribution is described by Coulomb term, and for van der Waals (vdW), the Lennard Jones term is measured (Figure 4.4).

The same trend is observed in both replicas - the van der Waals energy is the only contributor to the adsorption. In both replicas the vdW term converges to around -250 kJ mol^{-1}. In the EA10 this convergence happens at around 500 ns, but the -250 kJ mol^{-1} potential energy is already shown between 200 and 400 ns.

In the EA18 system, the energies do not converge until the rearrangement event takes place in the middle of the simulation. This is where the vdW term improves significantly from around -150 kJ mol^{-1} to -250 kJ mol^{-1} at around 550 ns time. Interestingly, the adsorption potential energy in the EA18 system appears stable already at 550 ns. This means that the change of residues which are involved in the adsorption at 720 ns have not been detected by the potential energy of adsorption.

Despite the different natures of residues close to the surface, the adsorption relies mostly on the hydrophobic interactions, or van der Walls interactions. This is explained by the composition of the FnIII domain. The hydrophobic core must be stabilised with the polar and charged residues on the surface. In the case of initial contact with any surface, it is expected that many of these residues will get trapped in the adsorption area. On the other hand, the functional groups in the surface are expected to be relatively well hydrated, particularly the partially charged esters in the EA SAMs, as measured in the previous chapter (Figure 3.5). The water at the surface likely helps to hydrate the polar and charged groups, suggesting a contribution of water-mediated adsorption. Unfortunately, this is cannot be captured with the potential energy of adsorption presented here.

4.4 RGD Motif Availability

The availability of the RGD motif for binding is probed by measuring whether the motif is closer to the surface than the 10^{th} domain (Figure 4.5). For both calculations the centre-of-mass of the domain was used, and for the surface the heavy atoms were used, as described in

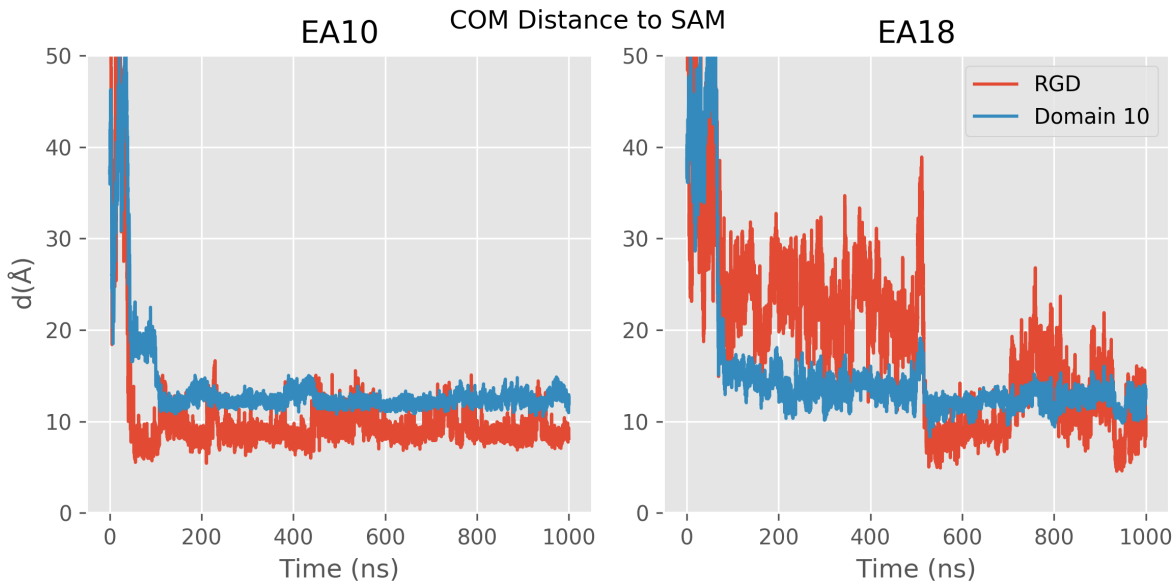


Figure 4.5: The minimum distance from the RGD motif (centre-of-mass), and the 10th domain (centre-of-mass) to the surface in both replicas. The motif is appears to reside on the side and is less accessible than the same motif in the tandem $FnIII_{9\text{-}10}$.

Chapter 3.

In the EA10 replica, the RGD motif is closer to the surface than the domain most of the time. The visual inspection of the simulation (not shown) shows that the motif is always placed to the side of the domain, placed perpendicularly to the surface, with Asp1495 touching the surface most of the time. This residue is also found to be close to the surface during the stably adsorbed periods (Table 4.1).

The behaviour of the motif is more complex in the EA18 system. A big transition takes place midway through the simulation. Initially, the domain is closer to the surface than the RGD, meaning that the motif is oriented towards the solvent and is available for binding. However, after the rearrangement event, the distance from the motif's centre-of-mass fluctuates between being closer to the surface and farther than the 10th domain. The visual inspection reveals that the motif is positioned in parallel to the surface, and even a small movement can change its centre-of-mass in the z dimension significantly. This happens despite the 10th domain being stably adsorbed. It is worth noting that the RGD availability changes significantly during the rearrangement event.

Whereas the motif cannot be easily buried in the interface considering how it is positioned

in the corner of the 10th domain, it can still be more or less available. Here, across the two replicas, the RGD availability is smaller than in the $FnIII_{9\text{-}10}$ tandem.

4.5 Secondary Structures

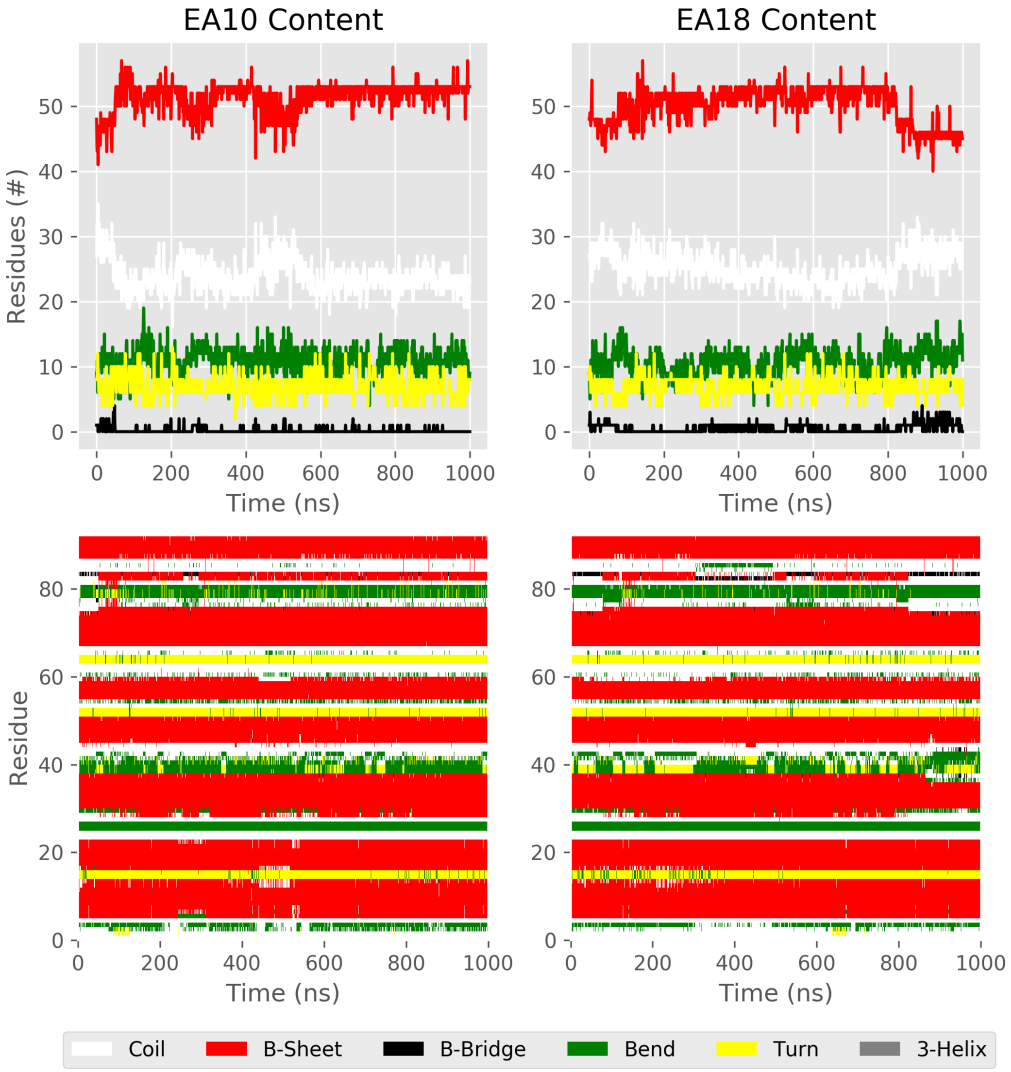


Figure 4.6: Secondary Structure Content in both EA_{10} replicas. The top two graphs represent the number of residues classified with any secondary structures which is computed with DSSP [2]. Note that the "Coil" represents unstructured loop areas - problem known with DSSP. The bottom two graphs are a close up view showing the evolution of the classification over time. The graphs on the **Left)** describe EA10 whereas on the **Right)** EA18 secondary structures are described. In the case of EA10, the secondary structure are fully conserved, whereas in the case of EA18, they are mostly conserved.

The secondary structure content of the 10th domain was measured with DSSP for both replicas [2] (Figure 4.6). In the case of the EA10 replica (upper-left) the number of residues assuming

the β-sheet structures oscillates at around 52. The more detailed picture in the bottom-left corner of the figure shows the DSSP classification over time for each residue. The classification of the residues are conserved for the duration of the simulation. The EA18 replica (upper-right) behaves similarly, with a small drop in the number of β-sheet starting from 800 ns time. This drop is spread across residues in the loop/turn areas between the different β - sheets (coloured red). However, it has not been observed in both replicas.

Overall, the secondary structures are conserved during adsorption. However, the relatively short timescale of the simulation limits the discussion of how the adsorption to the surface affects the secondary structures.

4.6 Discussion

The $FnIII_{10}$ domain adsorbs rapidly to the surface, which is in direct contrast to the adsorption of the domain in the $FnIII_{9\text{-}10}$ that was observed in the previous chapter. Therefore, it is concluded that the 9^{th} domain obstructs the adsorption of the 10^{th} domain in the tandem replicas. However, the potential energy of the adsorption in the tandem replicas does not outweigh clearly the energy of adsorption of the sole 10^{th} domain. One way to explain this is that the electrostatics of the 9^{th} domain make it more likely to initially lock in the two domains in the found orientation. This initial adsorption of the 9^{th} domain in the tandem replicas positions the 10^{th} domain in such a way that the relevant residues face away from the surface.

The RGD motif in the adsorption of the sole 10^{th} domain resides to the side of the protein. This orientation could make it less available than in the case of the $FnIII_{9\text{-}10}$ domains adsorption, where it is oriented towards the solvent. However, it is possible that removal of the 9^{th} domain makes the RGD motif more accessible. The motif accessibility cannot be directly translated to binding affinity with integrins, which is known to be better in the presence of the PHSRN motif on the 9^{th} domain.

The secondary structures of the 10^{th} domain during the adsorption are conserved throughout the EA10 replica, but a small decrease in the number of β-sheet is observed towards the end of

the EA18 replica. This replica is marked by the rearrangement event. This decrease could be used to justify extending the simulation - a problem inherent to molecular dynamics. How much sampling is enough? Which observable variables should be used to explain that the simulation has converged to a steady state? These questions are important and illuminate how important it is to quantify the convergence of simulations. However, in this thesis sampling is discussed via the use of multiple replicas. The mechanical stability of the 10^{th} has been probed via single-molecule force spectroscopy, along with other FnIII domains, showing that the FnIII_{10} domain is mechanically the weakest, which led the authors to suggest that stability is a factor in initiating fibrillogenesis [150]. Interestingly, weak mechanical stability was predicted a year earlier in steered molecular dynamics simulations [151]. Whereas the simulations discussed in this chapter have not shown any degradation of the secondary structures, the timespan of the simulations is likely to be too short for this kind of phenomena to be found. However, with the continuing progress in the cost of computing and sampling methods, the community might be able to probe this kind of questions in the near future.

In the previous chapter it was shown how a small difference in the surface chemistry affected the hydration of the EA and MA SAMs, and how that led to a very different adsorption profile of the tandem $\text{FnIII}_{9\text{-}10}$ domains. This adsorption was driven by the 9^{th} domain. Here it is shown that the 10^{th} domain when simulated by itself adsorbs promptly to the surface. The analysis of adsorption potential energies shows that in both chapters (3 and 4) the nature of the adsorption has an important hydrophobic component - relying largely on the van der Walls force with small or no contribution from Coulomb term. This complements the study in which the original polymer poly(ethyl acrylate) is modified to create two more hydrophobic polymers; on each of which Fn fibrillogenesis takes place [100]. However, the formed Fn networks have different patterns and more importantly, do not support cell differentiation as well as poly(ethyl acrylate). Therefore, in the next chapter we look into the adsorption of the $\text{FnIII}_{9\text{-}10}$ on a hydrophobic surface to see how it differs from the adsorption on the EA SAM and MA SAM surfaces.

Chapter 5

$FnIII_{9\text{-}10}$ on Methyl-functionalised SAMs

In the previous chapters the adsorption of $FnIII_{9\text{-}10}$ and $FnIII_{10}$ to the EA SAM was described. I showed that one extra methylene bridge in the EA SAM weakens the hydration of the surface, and leads to adsorption of both $FnIII_{9\text{-}10}$ and $FnIII_{10}$. The EA SAM was designed based on the polymer poly(ethyl acrylate), which has been recently compared further to its more hydrophobic variants poly(butyl acrylate) and poly(hexyl acrylate) [100]. Further addition of methylene bridges, which made the polymer more hydrophobic, led to a decline in cell differentiation. In this chapter I report the findings from my investigation of the adsorption of $FnIII_{9\text{-}10}$ to a model hydrophobic surface, methyl-terminated SAMs. During the analysis, the adsorption of the FnIII fragment to the surface and the decrease in cell differentiation in comparison to EA SAM is discussed.

5.1 Simulations

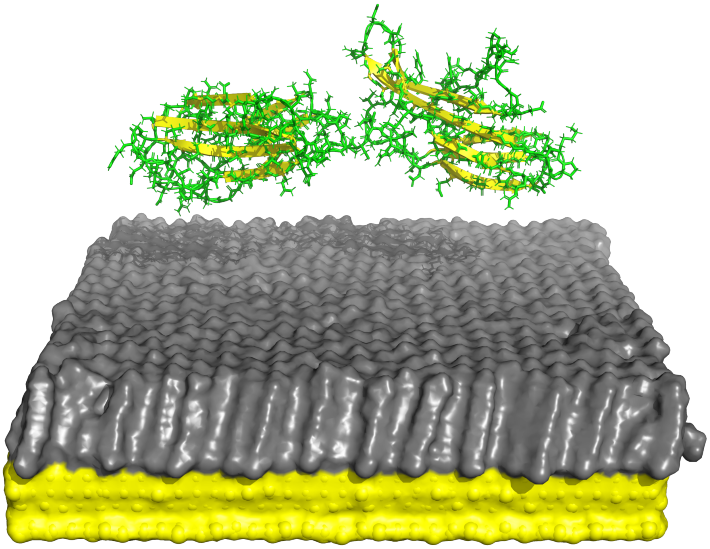


Figure 5.1: The system configuration with $FNIII_{9\text{-}10}$ placed on top of an equilibrated methyl SAMs. The water molecules and ions are not shown. The secondary structures are visualised as ribbon (yellow) and side chains as stick (green). The GOLD slab is at the bottom (yellow), whereas the methyl SAMs are coloured grey.

The surface was designed and equilibrated following the same protocol explained in Chapter 1 (3.1). The thiol self-assembled monolayer (SAM) coats a gold substrate, whereas the SAM functional groups are in contact with an aqueous environment. This environment contains the $FnIII_{9\text{-}10}$ peptide and the neutralising ions.

In order to mimic the previous design, two chain lengths $n = 10$ and 18 were used for the SAMs $SH(CH_2)_nCH_3$, which are used in the two different replicas. GROMACS version 5.x was used to simulate the two replicas, each of length 500 ns.The system was assembled, minimised and equilibrated in the same way the simulations in Chapter 3 (Figure 5.1).

The two SAMs with the different lengths converge to a similar tilt, with 49.33°for n = 10 and 48.29°for n = 18, which is why the two systems are referred to as replicas methyl10 and methyl18.

The interdomain orientation was quantified with a superdihedral which was calculated using the $C-\alpha$ atoms in the residues Ser1396-Val1345-Leu1434-Thr1486 (Figure 5.2). These four residues are in the centres of all 4 $\beta-Sheets$ across the two domains. The order of the residues

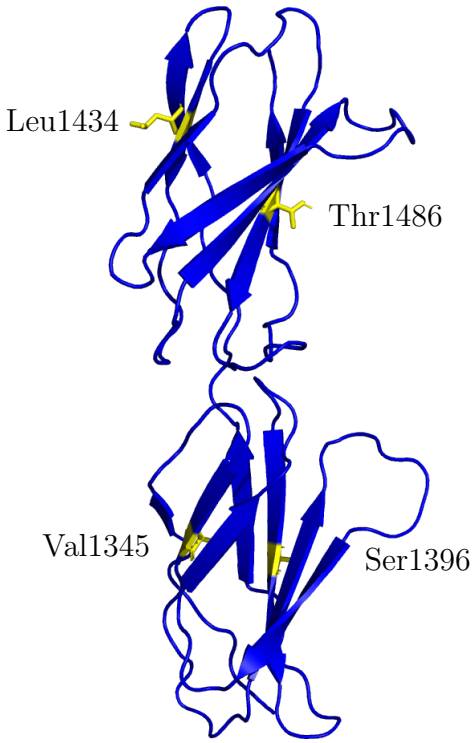


Figure 5.2: $FnIII_{9\text{-}10}$ domains with the four yellow-coloured labelled residues Ser1396-Val1345-Leu1434-Thr1486. The residues formed a part of the superdihedral used to quantify the rotation of one domain with respect to another.

was picked in such a way that the initial configuration is as close to 0° as possible. Any distances measured, unless explicitly stated, refer to the heavy atoms.

5.2 Adsorption

The adsorption was quantified by measuring the shortest distance from the centre-of-msas of each domain to the heavy atoms in the surface. In both methyl replicas the two domains $FnIII_{9\text{-}10}$ adsorb relatively quickly and stay stably at the surface for the remainder of the simulations (Figure 5.3).

Adsorption in the methyl10 system follows four distinctive stages (Figure 5.3, red and blue patches). The first three stages concern the adsorption of the 10^{th} domain. In the first stage, m10pI (2 - 123 ns), the 10^{th} domain makes initial contact. In the second stage, m10pII (124 - 162 ns), the domain shows an improved adsorption. In the third stage, m10pIII (162 - 500 ns), the domain is stably adsorbed as indicated by the lack of fluctuation in the COM-substrate

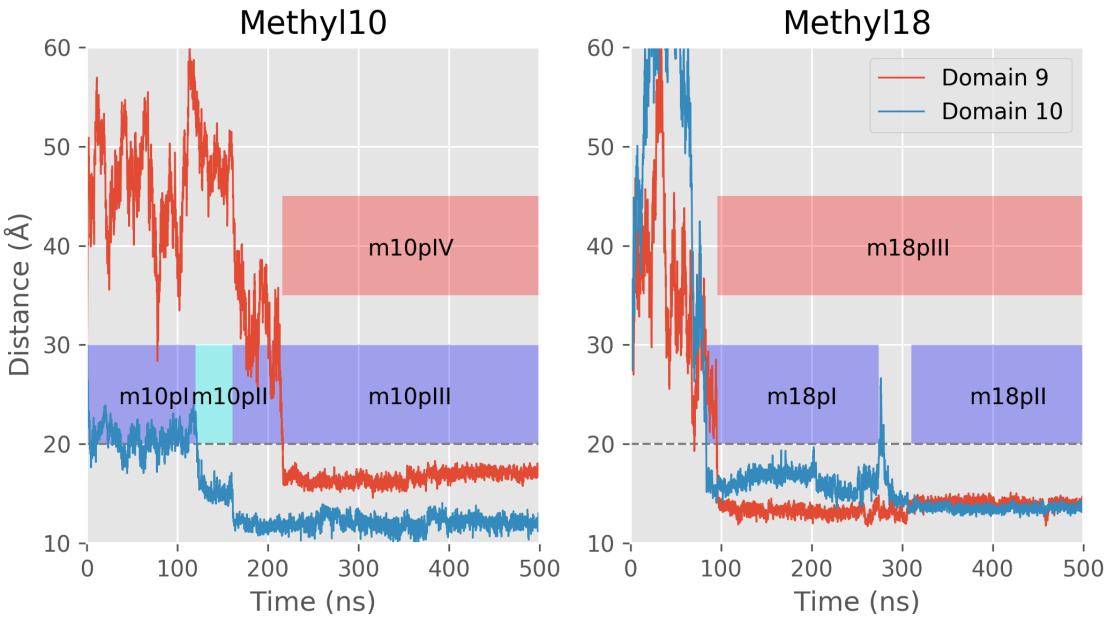


Figure 5.3: The distances from the centre-of-mass of domain 9 & 10 to the methyl SAMs surface for the two replicas methyl10 and methyl18. At a distance above 20 Å (grey dashed line), the domain is unlikely to be in contact with the substrate. The red- and blue-shaded patches highlight the different adhesion stages of the 9^{th} and 10^{th} domains, respectively.

distance. The last stage, m10pIV (217 - 500 ns), captures the sudden adsorption of the 9^{th} domain on the surface.

In the second replica, methyl18, three stages are defined in a similar fashion (coloured patches in Figure 5.3). The first two stages describe the adsorption of the 10^{th} domain: it slowly transitions to a more stable orientation during the first stage m18pI (84 - 274 ns), culminating in a rearrangement event in which the domain loses contact with the substrate for around 4 ns. After that the 10^{th} domain is stably adsorbed establishing the second stage m18pII (t = 310 - 500 ns). The third stage, m18pIII (96 - 500 ns), describes the relatively uniform adsorption of the 9^{th} domain. It is noted that midway through the simulation, the 9^{th} domain shows small fluctuations in its distance to the surface, which overlaps with the rearrangement event of the 10^{th} domain.

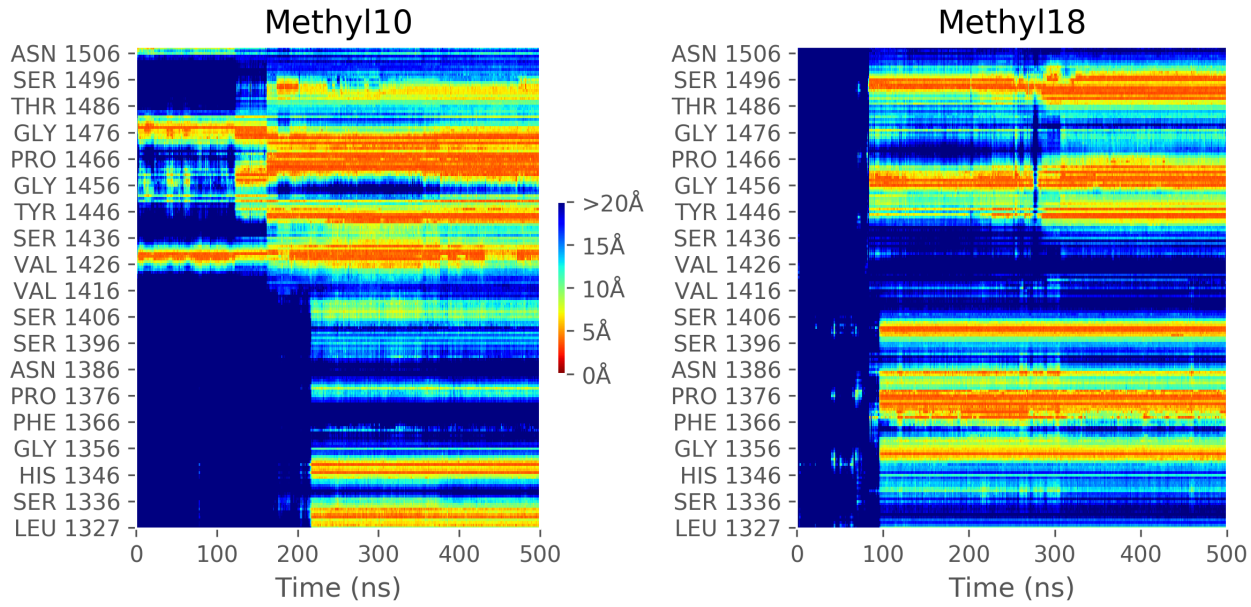


Figure 5.4: The distance maps visualise the nearest distance between the heavy atoms of each residue and the heavy atoms of the SAM. The data is presented for both methyl10 and methyl18 replicas separately.

5.2.1 Residue Adsorption

In this section a closer look is taken at the residues that drive the adsorption of the domains on the hydrophobic surface. The minimum distance is calculated from each residue to the SAM surface. These are presented together in Figure 5.4 in the form of distance maps that visualise the adsorption progress. The first thing that becomes clear by looking at the two contact maps for the methyl10 and methyl18 systems is that different residues are involved in the adsorption process.

In the methyl10 system, during the first stage m10pI, the 10^{th} domain touches the surface using two residue regions centred around Val1426 and Gly1476. This is followed by the shorter interim stage m10pII (124 - 162 ns), which involves two temporary regions around the residues Tyr1446 and Gly1456. In the third stage, m10pIII, the 10^{th} domain is stably adsorbed. This stability is clear from the continuity of the adsorbed residue regions. The regions which were involved in the interim stage disappear and give way to two strongly bound regions: one that spans Gly1456 to Gly1467, and one that involves a few residues around Tyr1446. In contrast, the 9^{th} domain shows very little variability in the way it adsorbs to the surface. It has mainly

	Interval (ns)	Domain 9	Interval (ns)	Domain 10
Methyl10	m10pIV: 217 - 500	Thr1331, Gly1332, Ile1348, Arg1351	m10pI: 2 - 123	Pro1430, Thr1431, Pro1479
			m10pII: 124 - 162	Thr1431, Pro1459, Gln1461, Ser1475 Gly1476, Leu1477, Lys1478, Pro1479
			m10pIII: 162 - 500	Thr1431, Ser1432, Arg1445, Tyr1446, Phe1463, Thr1464, Val1465 Pro1466, Gly1467, Ser1468, Lys1469, Thr1473, Ser1475, Gly1476
Methyl18	ma18pIII: 97 - 500	Thr1355, Asp1373 Arg1374, Val1375, Pro1376, His1377 Asn1401, Gly1402, Arg1403	m18pI: 84 - 274	Asn1457, Ser1458, Pro1459, Gly1494, Asp1495, Ser1496, Pro1497
			m18pII: 310 - 500	Arg1445, Tyr1446, Arg1448, Asn1457, Ser1458, Pro1459, Val1460, Gln1461, Glu1462 Thr1464, Val1490, Thr1491, Gly1492, Arg1493, Gly1494, Asp1495, Pro1497, Ala1498

Table 5.1: Residues that are less than 6 Å away from the surface for at least 80% of the adsorption stage for the hydrophobic Methyl SAMs.

two residue regions which are found in the first 25 residues of the N-terminal.

The adsorption of the two $FnIII_{9\text{-}10}$ in the methyl18 system consists of fewer events. Both domains adsorb within 100 ns of the simulation. The lower 89 residues in the graph represent the adsorption of the 9^{th} domain, which does not change throughout the simulation (stage m18pIII). However, the top 94 residues in the distance map describe the adsorption of the 10^{th} domain, show a change in adsorption midway. In the 10^{th} domain, two residue regions are adsorbed stably initially (stage m18pI). Midway through the simulation, between 274 and 310 ns, the rearrangement event takes place - the residues in the Gly1456 area temporarily move away from the surface. After the rearrangement, the residues come back to the surface and more adjacent residues make contact (m18pII). Furthermore, during this stage another residue region surrounding Tyr1446 adsorbs to the surface.

For each adsorption stage the residues residing within 6 Å of the surface are listed in Table 5.1. Across the two replicas, the 10^{th} domain adsorbs via a variety of different residues. The only exceptions are Pro1459 (m10pII and m18pI) and three common residues found in the last adsorption stages (Arg1445, Tyr1446, Thr1464, circled in Figure 5.5).

In the methyl10 replica, from stage m10pI to m10pII, several new residues come close to the SAMs interface: the hydrophobic Pro1459, Leu1477, a few polar residues and the charged residue Lys1478. Furthermore, the last stage m10pIII involves the hydrophobic residues Phe1463 and Val1465, more polar residues and the charged residues Arg1445 and Lys1469. In the methyl18 system, from ma18pI to ma18pII stage, only two additional hydrophobic residues come close to the surface (Tyr1446, Val1490), and many charged residues (Arg1445, Arg1448, Glu1462, Arg1493, Gly1494).

Several charged and polar residues reside close to the methyl18 SAMs substrate. These residues

seem to be oriented such that the charged side chains lie parallel to the interface with the SAMs directed away from the protein. Such orientation would allow the side chains to maximise their hydration (not shown).

On both methyl substrates the two domains adsorb quickly and stay adsorbed for the remainder of the simulations. However, despite having as many as 25 residues interacting with the substrate during the stably adsorbed stages, only a few residues are common across the two replicas (Table 5.1). This lack of binding specificity is due to the hydrophobicity of the methyl SAMs and the many hydrophobic patches present on the domains 9 and 10.

5.3 Surface Electrostatics

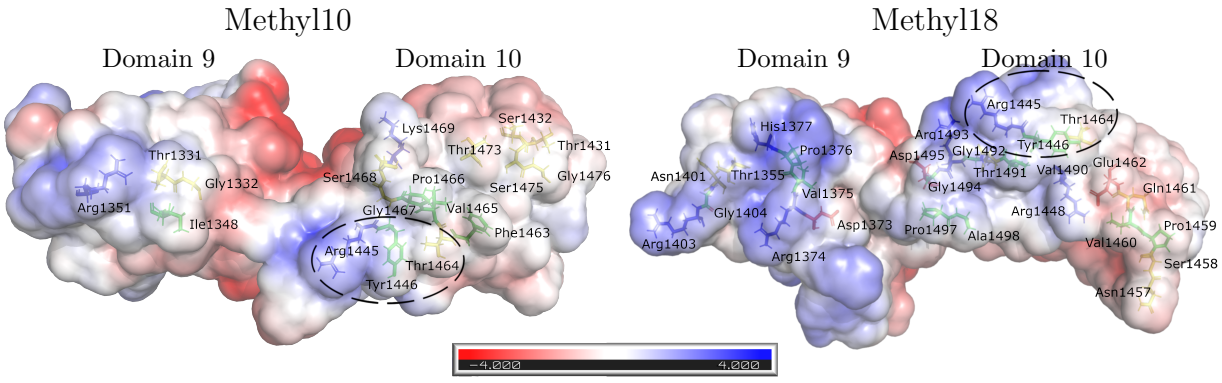


Figure 5.5: The electrostatic surface potential of the final adsorption state in methyl10 (**left**) and in methyl18 (**right**) with annotated residues which were close to the substrate during the final adsorption stages (see Table 5.1). The three residues enclosed in the dashed ellipses are common between the methyl10 and methyl18 replicas.

The electrostatic surface potential of the $FnIII_{9\text{-}10}$ is presented in Figure 5.5 showing the adsorbed faces of the protein fragment on methyl10 and methyl18. In the methyl10 replica the adsorbed protein face contains a large circular hydrophobic patch (white colour) on the 10^{th} domain. This patch involves Tyr1446, Phe1463, Val1465 and Pro1466 at the centre of the interactions. In contrast, the 9^{th} domain has a small hydrophobic patch comprising Gly1332 and Ile1348. The methyl18 system presents a more complicated picture. The 10^{th} domain interacts with a different face as indicated by the small number of common residues Arg1445, Tyr1446 and Thr1464 (dashed ellipse). The domain presents an elongated hydrophobic path, starting from Pro1497, Ala1498 through Val1490 and Tyr1446 to Val1460 and Pro1459. The

9th domain, surprisingly, presents a slightly positive face to the hydrophobic surface, with only two hydrophobic residues Val1375 and Pro1376.

5.4 Potential Interaction Energy

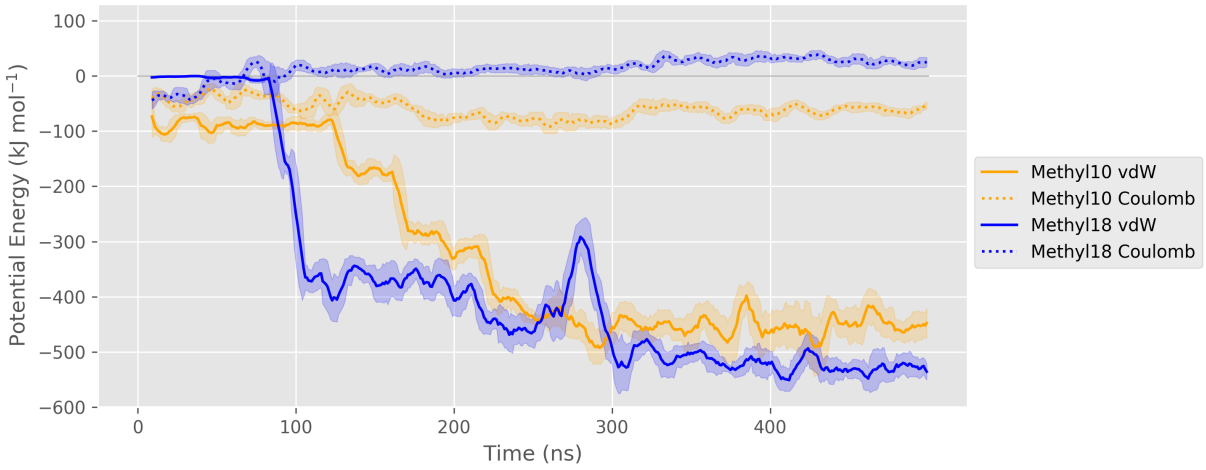


Figure 5.6: The non-bonded potential energy terms between the protein and the methyl SAMs. The van der Waals (vdW) interactions are described by the Lennard Jones potential, whereas the electrostatics is described by the Coulomb potential. The rolling mean (dashed-line) and standard deviation (coloured-area) is used with the window size of 1 ns.

The interactions between the methyl substrates and the domains are solely due to the van der Waals contribution (Figure 5.6). On methyl10, a decrease in the van der Waals potential energy at 120 ns is observed, which continues until 250 ns through all adsorption stages of the 9th and 10th domains. On methyl18 the protein shows a more drastic change in van der Waals term (solid line) when the two domains adsorb suddenly. After the rearrangement of the 10th domain observed just before t = 300 ns, a steady decrease of the potential energy to slightly below -500 kJ/mol is observed, indicating that even during the most stable adsorption stage m18pII subtle changes continue taking place.

The potential energies of the interaction with the hydrophobic substrates are significantly more favourable than with the EA SAMs 3.2.2, indicating preferential adsorption of FnIII $_{9\text{-}10}$ onto hydrophobic surfaces.

5.5 Structural Motifs RGD and PHSRN

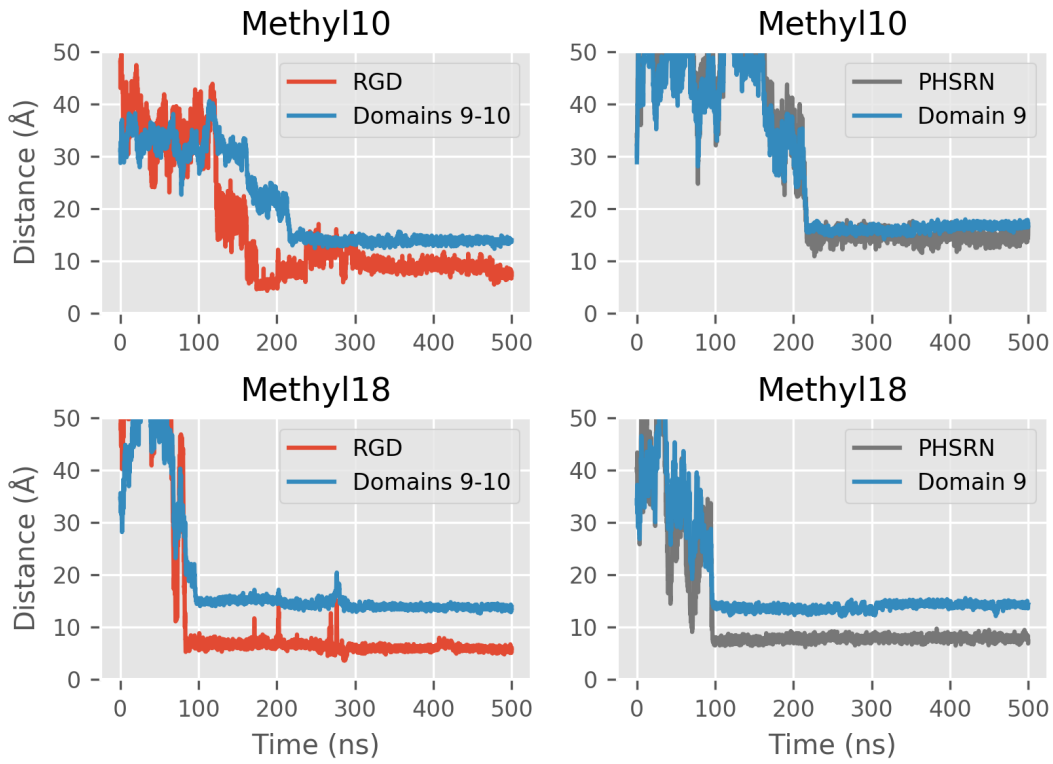


Figure 5.7: RGD (left) and PHSRN (right) motif exposure to potential interactions with integrin receptors. The graphs present the distances from the centre-of-mass of a motif, and of a protein fragment, to the surface. For the PHSRN motif, the centre-of-mass of the 9^{th} was used, and for RGD motif, the centre-of-mass of both domains was used.

The motifs RGD and PHSRN are important for binding to receptors from the integrin family. To understand how exposed these motifs are, we check whether they are closer to the substrate than the protein fragments to which they are attached (Figure 5.7). A motif being closer to the surface suggests that it is either buried in between the surface and the protein, or that the protein adsorbs in a way that partly makes the motif inaccessible.

The RGD motif is significantly less accessible on the methyl SAMs than on EA SAMs (Chapter 3). The methyl18 system is perhaps simplest to describe: the RGD motif is significantly closer to the surface than the centre-of-mass of the two domains. This behaviour continues throughout the adsorbed period. Visual inspection of the simulation shows that the motif is trapped in between the protein and the surface. The behaviour of RGD on methyl10 follows a more complex pattern. At around 250 ns the RGD motif is as far away from the surface as

the centre-of-mass of the two domains. However, this is followed by a slow transition where the RGD moves closer to the surface, becoming less accessible. The overall motif distance to the surface is below 10 Å during the adsorbed period. Similarly, visual inspection confirms that the motif is trapped between the protein and the surface, which is consistent with the methyl18 replica.

The PHSRN motif is also closer to the substrate than the centre-of-mass of the $FnIII_9$ domain in the methyl18 system. In the methyl10, the motif is more or less the same distance away from the substrate as the centre-of-mass of the 9^{th} domain (around 13 Å). This suggests that the motif is on the side of the domain, which is confirmed visually (not shown). The methyl18 replica shows that the motif is also buried in the surface. However, watching the simulation reveals that the motif is partly buried in the surface and partly accessible to the side of the 9^{th} domain which is consistent across both systems. This burial of motifs on the methyl systems is in stark contrast to the EA systems which frequently orient the motifs towards the solvent. Both motifs across the EA systems are found farther away from the interface (Figure 3.6).

Interestingly, the exposure of RGD and PHSRN in the methyl10 system appear to diverge. RGD becomes increasingly buried while PHSRN stays in the same position. However, the two motifs are present on the same face of the protein. This suggests that the way the two domains are aligned undergoes significant changes - changes which take place while the protein is adsorbed.

5.6 Interdomain Orientation

The divergence in the availability of motifs exposes that the interdomain relationship in the $FnIII_{9\text{-}10}$ changes. In this section, this relationship is quantified by measuring two quantities. First is an angle between the 9^{th} domain, the linker, and 10^{th} domain. Second is a super-dihedral angle that describes the rotation of the domains with respect to each other (Figure 5.8). The super-dihedral angle is calculated using the four residues Ser1396-Val1345-Leu1434-Thr1486 which represent the four $\beta - Sheets$ across the two domains (see 5.1).

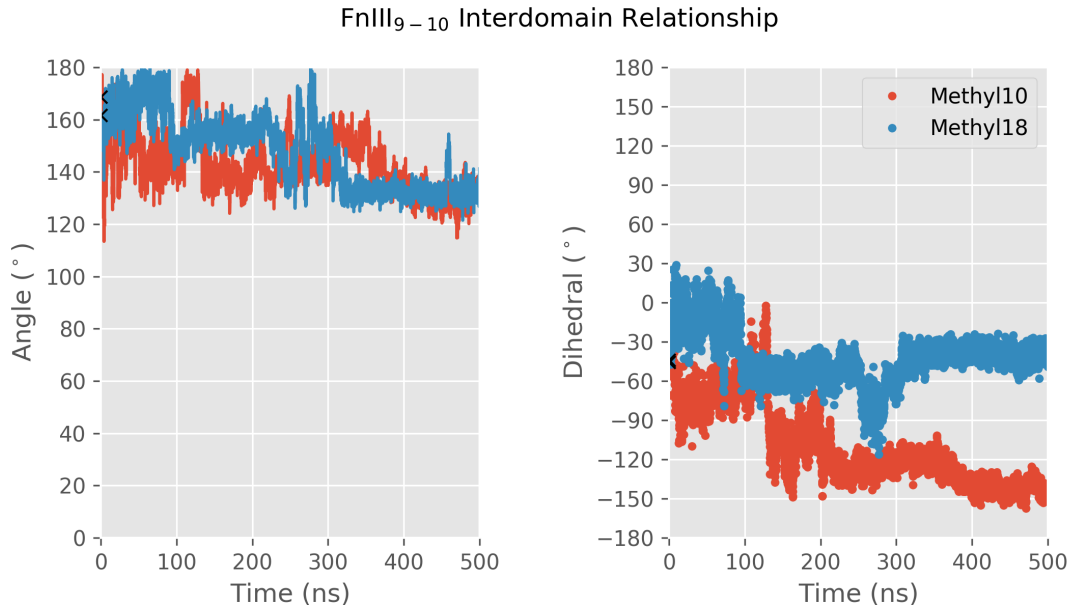


Figure 5.8: The description of interdomain orientation for the 9th and 10th domain during the adsorption. The angle was calculated between the centre-of-mass of the 9th domain, the linker, and the 10th domain. For the dihedral angle four anchor residues were used which comprise the β - strand secondary structures as described in methods. The black crosses X represent initial angles and dihedrals.

The 9th and 10th domains have an interdomain angle of around 165°. Towards the end of simulations, there is a small bend observed in both replicas, with the angle stabilising to around 130°. The change in the angle from around 165° to 130° takes place when the protein is adsorbed - after 300 ns time. During this time the adsorption potential energy on these hydrophobic surfaces is strong, which could force the protein into this new interdomain angle, which is a departure from the crystal structure (165°).

The interdomain rotation, or twist, was quantified with a super-dihedral that is based on four residues with each being embedded into a different β-sheet across the two domains (Section 5.1). Across the two replicas the initial dihedral is set to around -45°(Figure 5.8). In the methyl18 system, the dihedral at the end oscillates around the initial starting value. In contrast, in the methyl10 system, the dihedral diverges to -140°, a value around which it stabilises towards the end of the adsorption. This means that the two domains have rotated more than 90° with respect to one another.

This rotation shows why the exposure of RGD and PHSRN diverge. It is noted that, with respect to the surface, it is the 10th domain that rotates.

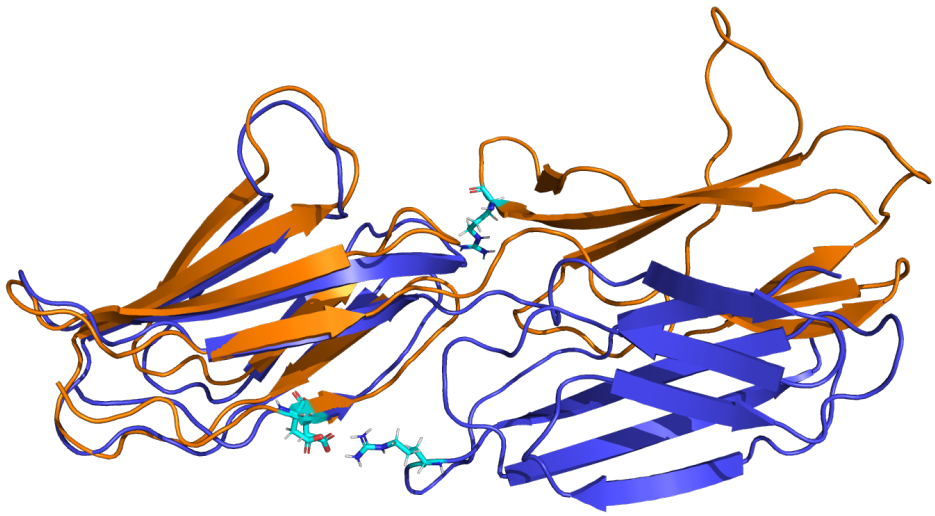


Figure 5.9: The initial (orange) and final state (blue) of the $FnIII_{9\text{-}10}$ in the methyl10 system. The domains were superimposed using the 9^{th} domain. The 10^{th} domain (right) rotates with respect to the 9^{th} domain (left). Arg1493 in the RGD motif and the aspartic acid Asp1334 form two hydrogen bonds with each other and are visualised using the licorice representation. The two residues are buried in the surface.

The change in the interdomain orientation in the methyl10 system is visualised in Figure 5.9. The C_α atoms in the 9^{th} domain were superimposed onto each other using the protein at the first and last frame. Arg1493 is highlighted in the 10^{th} domain together with the negatively charged Asp1334 in the 9^{th} domain. This visualises the significant rotation of the 10^{th} domain with respect to the 9^{th} domain, which took place during the adsorption. The two highlighted residues form two hydrogen bonds each, which likely partly stabilises this interdomain rotation.

The observed change in the interdomain orientation affects the exposure of the RGD and PHSRN motifs. After the interdomain change, the two motifs reside on different sides of the protein, which changes the relative distance between them. For this reason the ability of the two domains to bind to integrin receptors might be adversely affected. This is because the RGD-PHSRN might be playing an important role in this binding [152].

5.7 Discussion

In the two methyl replicas both domains $FnIII_{9\text{-}10}$ adsorb well to the surface. In the process, different residues are used, highlighting the characteristic lack of adsorption specificity on hydrophobic surfaces The adsorption relies solely on van der Waals force. Furthermore, the electrostatic surface potential shows that the hydrophobic patches and paths on the adsorbed faces of the protein are used during the adsorption.

This lack of adsorption specificity has been previously observed in the simulations of the 9^{th} domain on a gold slab in [94]. Another set of simulations by the same group also showed that adsorption of $FnIII_{8\text{-}10}$ methyl-terminated SAMs is non-specific [93].

In the process of adsorption the RGD and PHSRN motifs are shown to be buried in the surface, which would stop any possible binding to integrins. This could explain why osteoblast-like cells on the methyl SAMs coated with $FnIII_{7\text{-}10}$ do not form actin cytoskeleton or focal adhesion sites, in contrast to the polymer poly(ethyl acrylate) [153]. This is also consistent with the finding in Chapter 3 where the two motifs are shown to be exposed for binding on the EA SAM, which models the polymer poly(ethyl acrylate).

Denaturation The findings of this chapter and observations made during the analysis of these simulations have also provided insight into the denaturation of the domains on the hydrophobic interface. During the analysis of residues on the hydrophobic surface it was noticed that several charged residues are found close to the surface (5.1). For example the charged residues Arg1351, Arg1445 and Lys1469 are close to the surface during the adsorbed stage of the methyl10 system. In the methyl18 system, three residues in the 9^{th} domain, and five charged residues in the 10^{th} domain are listed. These residues happen to be close to the surface which does not necessarily mean that they mitigate the adsorption. While watching how the charged residues behave in the two simulations it appears that they mostly have their side chains reaching out to the sides. It seems that they try to stay solvated rather than being trapped between the protein (hydrophobic core) and the hydrophobic surface. This suggests a plausible mechanism for the denaturation of the FnIII type domain - the hydrophilic residues trapped between the

hydrophobic core and the hydrophobic surface try to stay solvated. In the process, they open up the hydrophobic core to the methyl substrate, leading to the loss of a tertiary structure. This could explain why our collaborators Annie Zhe Cheng from the group of Prof. Manuel Salmeron-Sanchez at the University of Glasgow, could not detect fibronectin with polyclonal antibodies on methyl SAMs experimentally after they formed Fn network, in contrast to EA SAMs (unpublished).

In the methyl10 system, a different interdomain orientation is assumed by the two domains, which is likely due to the van der Waals interactions on the hydrophobic surface. How significant are the interdomain rotation and bend? The interdomain interface between $FnIII_9$ and $FnIII_{10}$ covers around 300 Å^2 [1]. This is almost half of the estimated 550 Å^2 buried in the $FnIII_{7/8}$ and $FnIII_{8/9}$ interfaces [1]. The smaller buried interface $FnIII_{9/10}$ mean two things. Firstly, more residues are accessible for contacts, and secondly, the interdomain conformation between the two domains is less rigid. The latter is observed in the methyl10 system. However, this interdomain rotation was not observed in Chapter 3 where the $FnIII_{9\text{-}10}$ adsorbed on EA SAMs. Instead, the 9$^{\text{th}}$ domain inhibited the adsorption of the 10$^{\text{th}}$ domain, as shown in Chapter 4 where the 10$^{\text{th}}$ adsorbed well by itself. This leads to the question, is there a $FnIII_{9\text{-}10}$ orientation such that the two domains can both adsorb to the EA surface? Is the smaller buried interface between the two domains of importance? In the next two chapters, a closer look is taken at how the CHARMM36 and CHARMM36m forcefields affect the $FnIII_{9\text{-}10}$ complex, which is followed by the analysis of the rotation of 10$^{\text{th}}$ domain with respect to the 9$^{\text{th}}$ in bulk simulations.

Chapter 6

The Instability of $FnIII_{9\text{-}10}$ with CHARMM36

In the previous chapter it was shown that the $FnIII_{9\text{-}10}$ adsorbed promptly to the hydrophobic self-assembled monolayers (SAMs). In one of the replicas, the interdomain interface undergoes substantial changes, with one domain rotating over 90° with respect to another. Despite this apparent ability of the two domains to bend and rotate, it was concluded from Chapters 3 and 4 that the 9^{th} domain stops the 10^{th} domain from adsorbing on the EA SAMs, and that the 10^{th} domain adsorbs very well by itself. Is there an orientation such that both the 9^{th} and 10^{th} domains can adsorb to the surface? However, before answering this question, the impact of the CHARMM36 and CHARMM36m forcefields on the $FnIII_{9\text{-}10}$ complex must be discussed. This discussion is necessary due to finding unexpected instability of tertiary structure in bulk water with CHARMM36, which could affect the analysis of $FnIII_{9\text{-}10}$ interdomain orientation. The updated version of CHARMM36, CHARMM36m [120], is parametrised using intrinsically disordered proteins in addition to the previous datasets. In this chapter the impact of CHARMM36 and CHARMM36m on the $FnIII_{9\text{-}10}$ complex in bulk water is analysed.

6.1 Simulations

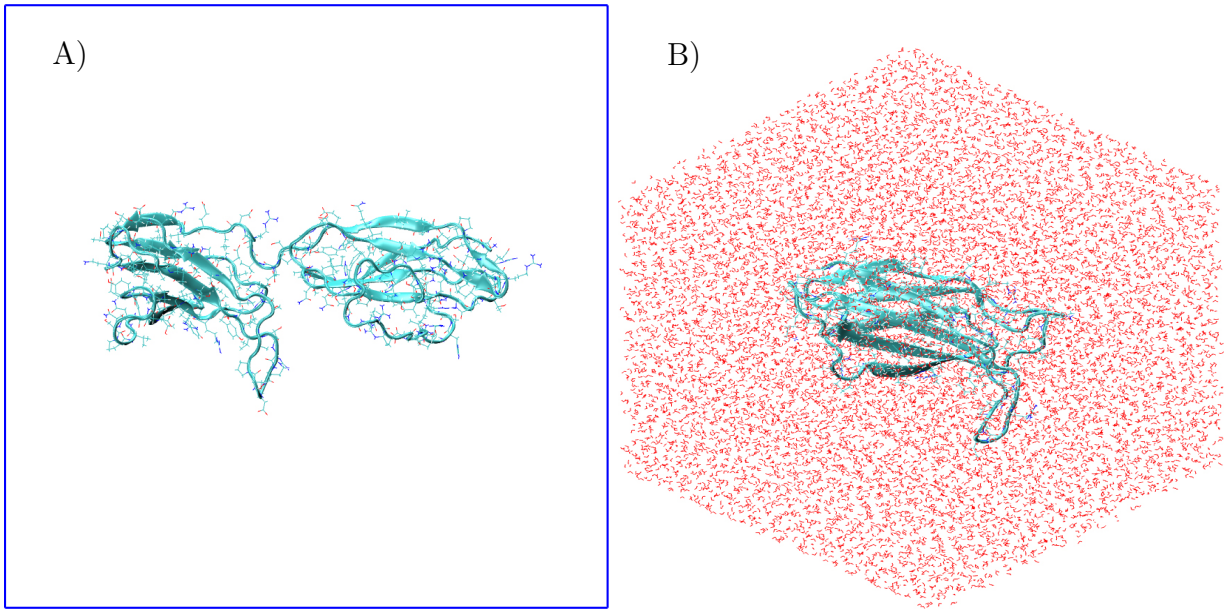


Figure 6.1: Initial system configuration examples. The secondary structures are shown as ribbon (cyan) and side chains as lines (colour is type dependant). A) Neutralised $FNIII_{9\text{-}10}$ in bulk water. The water molecules and ions are not shown. The PBC box is coloured blue. B) A system with a single domain ($FNIII_{10}$). The water molecules shown visualise the dodecahedron PBC box.

Altogether, eight simulations were carried out to investigate the effect of CHARMM36 and CHARMM36m. The first pair of simulations contain the two domains $FnIII_{9\text{-}10}$ in solution and use the CHARMM36 forcefield. A triclinic box was used for the periodic boundary condition with at least 20 Å between any atom of the protein and the boundary (Figure 6.1A). Thus, the box was around 109 Å long in every dimension. After the addition of water and neutralising atoms the system was minimised to a maximum potential force of 500 kJ mol^{-1} nm^{-1}. Then, the system was equilibrated for 100 ps in the NVT ensemble using the Nosé-Hoover at a temperature of 300K. LINCS was applied to all hydrogen bonds. PME was used as described in Chapter 3. The two production replicas were simulated in the NPT ensemble with each being 500 ns long at a temperature of 300K. The isotropic Parrinello-Rahman barostat was used to maintain the pressure at 1 bar. Each system contained almost 130,000 atoms.

Four single-domain simulations were simulated: two for the 9^{th} domain and two for the 10^{th} domain. The same protocol outlined in the previous paragraph was followed except for the box type, which was changed to a dodecahedron (Figure 6.1B). The smaller protein along with the

change in box type led to a system with only 40,000 atoms.

The last set of simulations include two replicas, each 500 ns long, which use the CHARMM36m and contain the FnIII$_{9\text{-}10}$ complex. The systems comprised over 124,000 atoms.

In this chapter, the structural RMS deviation is calculated with respect to the energy minimised structure. This is done in order to highlight the deviation from the minimised structure over the time of the trajectory, rather than compare it to the crystal structure where the RGD loop can create a false signal.

6.2 Stability of FnIII$_{9\text{-}10}$ with CHARMM36

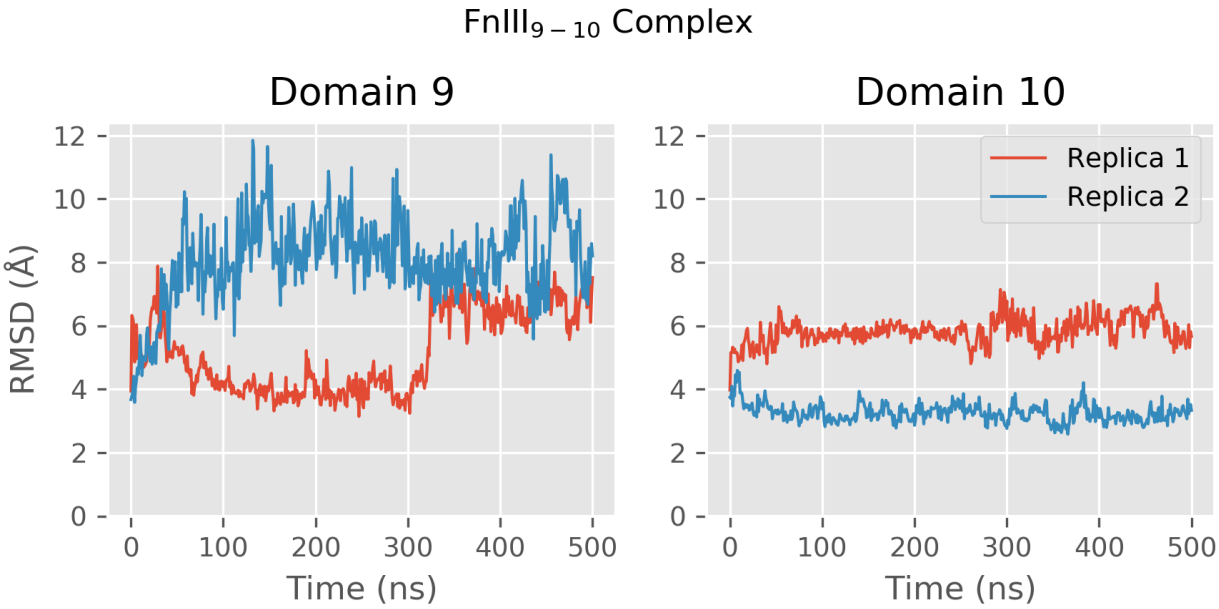


Figure 6.2: The RMSD of each domain in the tandem simulations FnIII$_{9\text{-}10}$. The value was calculated separately for the 9$^{\text{th}}$ and 10$^{\text{th}}$ domain. The minimised structure of each domain was used as a reference point.

Two replica systems containing the two FnIII$_{9\text{-}10}$ domains were simulated for 500 ns. The structural stability is described by calculating the root mean square deviation (RMSD) with respect to the initial structure that had its potential energy minimised. The RMSD was calculated for each domain in each replica, as shown in Figure 6.2. In both cases the 9$^{\text{th}}$ domain diverges from the original structure, although this happens to a larger extent in the second replica. The 10$^{\text{th}}$ domain, on the other hand, is more conserved in the second replica.

The RMSD of the 9th domain in the second replica increases shortly after the start of the simulation, frequently having a 10 Å RMSD distance from the reference structure. This large difference indicates that the domain undergoes some form of structural break down. In the first replica, during the first 40 ns the domain appears to be losing structure. The domain then remains stable until around 320 ns, with the RMSD oscillating around 4Å, which is relatively close to the original structure. However, after that a sudden increase in the RMSD to above 6 Å takes place. This value is similar to the RMSD seen in the second replica.

The RMSD of the 10th domain does not diverge to a similar extent. This is the case particularly in the second replica, where the RMSD oscillates steadily at around 4 Å. However, in the first replica, the RMSD increases to 6 Å and oscillates around this value for the rest of the simulation.

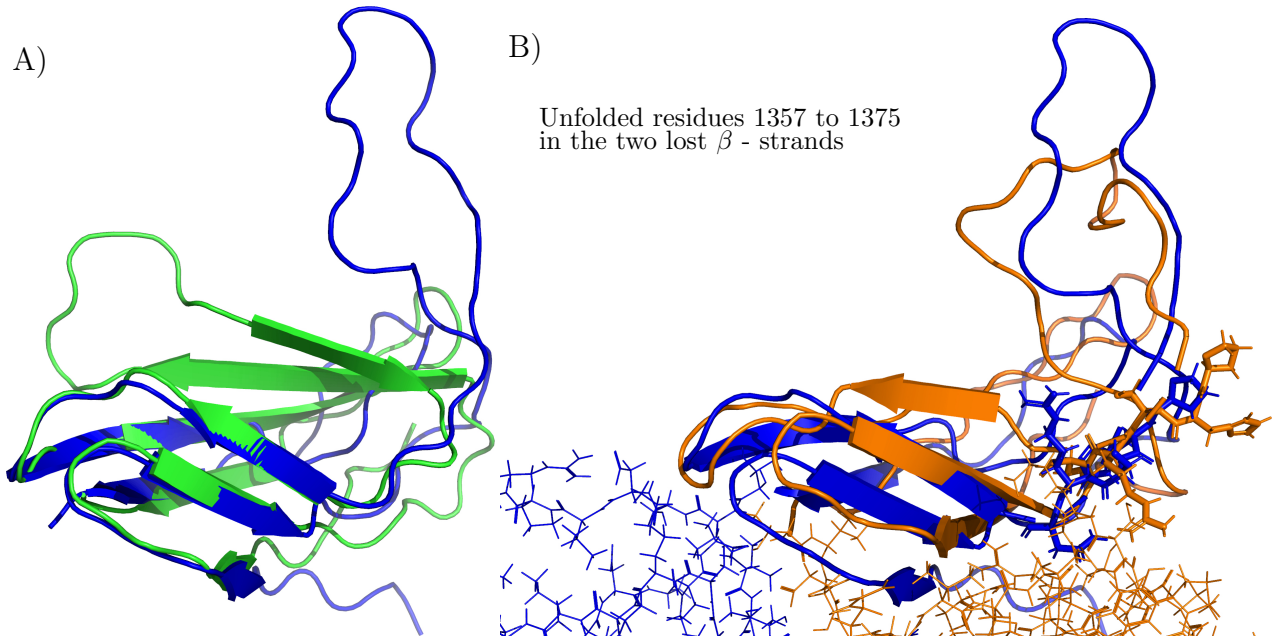


Figure 6.3: The denaturation of the 9th domain. The partly unfolded 9th domain in $FnIII_{9\text{-}10}$ at the end of the first (orange) and the second (blue) replica illuminate the meaning behind the large RMSD value of the 9th domain (Figure 6.2). A) The denatured 9th domain from the second replica is superimposed with the energy minimised 9th crystal structure (green) to highlight the loss of the two β - strands. B) The denatured 9th domain structures were superimposed. The synergy region uses licorice representation, whereas the 10th domain is shown as lines.

In Figure 6.3, the last states of the 9th domain are presented for each replica. In both cases two loops unfold and appear detached from the β - sandwich tertiary structure. The loops were previously β - strands. The unfolded region contains residues 1357 to 1375, which are followed by the PHSRN motif. These large changes show clear partial unfolding of the tertiary structure

of the 9th domain. Interestingly, in stark contrast, there is no unfolding seen of the 10th domain (not shown), despite the large RMSD value seen in the first replica.

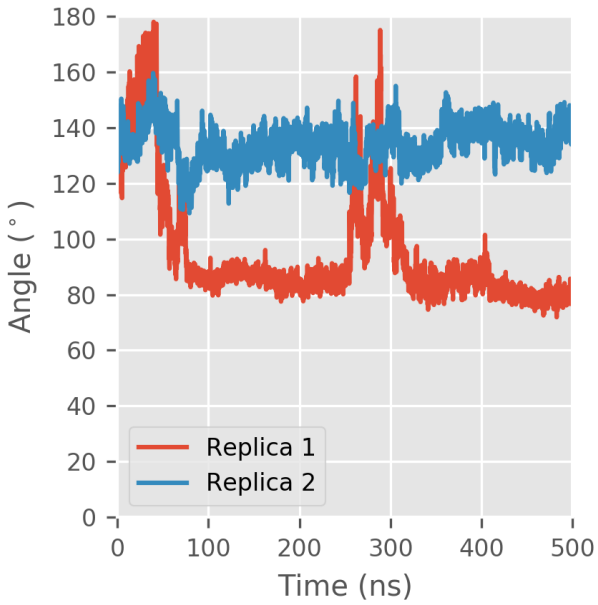


Figure 6.4: The angle between the centre-of-masses of the 9th domain, the linker, and the 10th domain (see Methods) for the two replicas.

The partial unfolding of the 9th domain is unexpected [154, 155]. In both replicas the FnIII$_{9\text{-}10}$ domains have been observed to bend and interact with each other. As was shown in Figure 5.9, a bend between the 9th domain and the 10th domain has been observed during the adsorption to the hydrophobic methyl-terminated SAMs surface.

The analysis of the angle between the two domains reveals that in the first replica, the two domains are strongly bent, with the angle reaching 80°. This is a small angle which is a rare occurrence in murine fibronectin [155]. In the second replica the angle oscillates at around 130°, which is also a small angle in comparison to the starting position. It should be pointed out that the calculation relies on the centre-of-mass of the 9th domain, which is affected by the partial unfolding. However, the angle between the domains appears to be decreasing even in the beginning of the two simulations. Therefore, the bent configuration likely proceeds the unfolding of the 9th domain.

In these two replicas, the FnIII$_{9\text{-}10}$ domains bend and interact with each other, preceding the partial unfolding of the 9th domain. Judging by the angle between the two domains, the 10th

domain interacts differently with the 9th domain in each replica. And yet, the 9th domain unfolds in the same way in each replica.

6.3 Stability of Lone FnIII$_9$ and FnIII$_{10}$

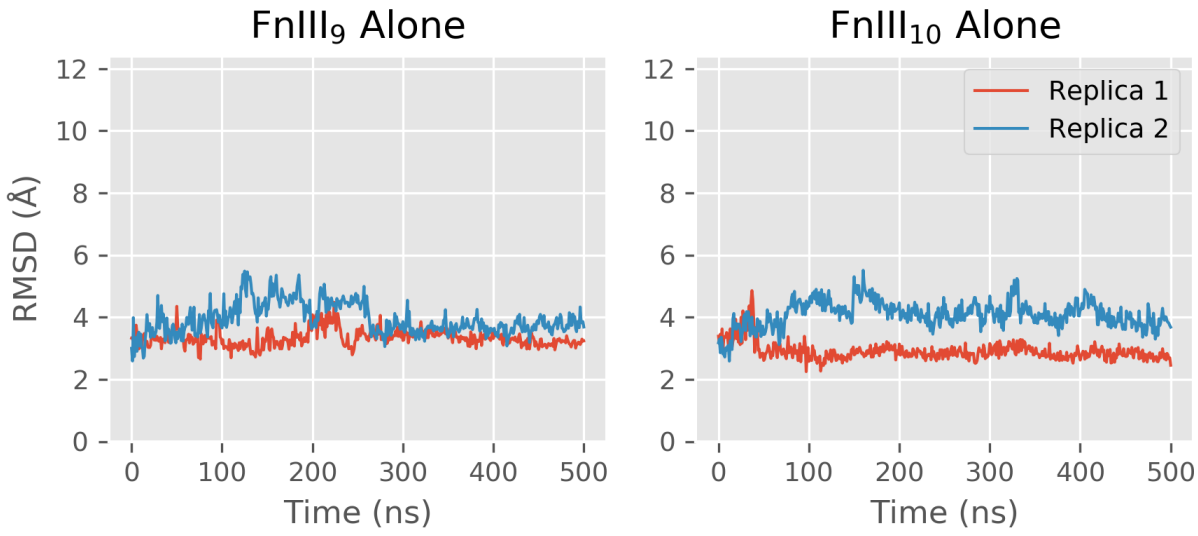


Figure 6.5: The RMSD of the FnIII$_9$ and FnIII$_{10}$ domains simulated separately. Both domains remain similar to the starting structure throughout the simulation, with the 10th domain showing small fluctuations.

In order to test whether the interactions between the 9th and the 10th domain affect their stability, two replicas of each were simulated. The first pair contains only the 9th domain, whereas the second pair contains only the 10th domain. The RMSD is calculated and shown in Figure 6.5 for each domain and replica. Both domains, when simulated alone, are significantly more stable than when simulated in FnIII$_{9\text{-}10}$ complex.

The RMSD of the 9th domain simulated alone remains mostly below 4 Å, with little indication of instability or unfolding. This is in large contrast to the instability of the domain seen in FnIII$_{9\text{-}10}$ complex, where it unfolds in each replica (Figure 6.3). The RMSD of the 10th domain simulated alone is also smaller, with a very small value seen in the first replica. Whereas the 10th domain did not unfold in FnIII$_{9\text{-}10}$ complex, in one replica, its RMSD increased to around 6 Å. In the absence of the 9th domain, the RMSD of the 10th domain converges to around 3

and 4 Å RMSD. These are on average smaller changes than are found in the FnIII$_{9\text{-}10}$ complex.

Therefore, it is concluded that the interactions between the 9$^{\text{th}}$ and 10$^{\text{th}}$ domain lead to partial unfolding of the 9$^{\text{th}}$ domain, and to an increase in RMSD of the 10$^{\text{th}}$ domain.

6.3.1 Cross-PBC interactions

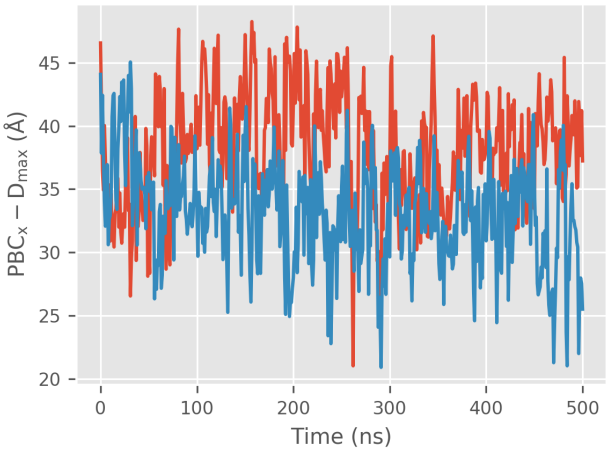


Figure 6.6: The closest possible distance between the protein and its PBC image. The distance is never less than 20 Å. PBC$_x$ is the length of the PBC in x dimension, and D$_{max}$ is the distance between the farthest atoms in the protein.

The original size of the system was made large enough to ensure that the protein does not interact with itself across the PBC. However, the denaturation of the protein increases its size. Here I check whether the denaturation and the observed states of the protein are not an artefact of interactions across the PBC. To do this, for each frame, I calculated the PBC$_x$ − D$_{max}$, where PBC$_x$ is the length of the x dimension in the PBC (the others are the same because the isotropic NPT was used), and D$_{max}$ is the distance between the farthest atoms in the protein. The results (Figure 6.6) show clearly that there is always more than 20 Å distance between the protein and its PBC image. Furthermore, the protein is unlikely to be aligned with any of the PBC vectors. Therefore, the protein does not interact directly with itself through the PBC box.

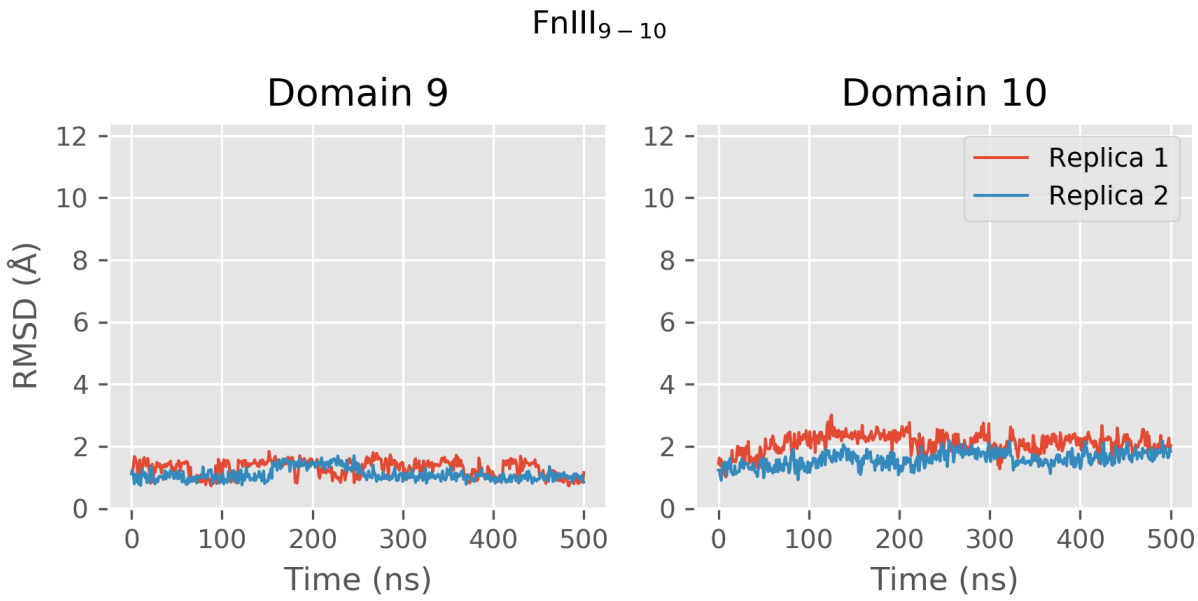


Figure 6.7: The RMSD of each domain measured separately in the FnIII$_{9\text{-}10}$ tandem and the udpated forcefield CHARMM36m. Both domains remain surprisingly stable throughout the simulation. The RMSD was measured with respect to the energy minimised structure with the same forcefield.

6.4 Stability of FnIII$_{9\text{-}10}$ with CHARMM36m

The FnIII$_{9\text{-}10}$ domains are very stable when simulated with a CHARMM36m forcefield (Figure 6.7). The 9$^{\text{th}}$ domain is never further than 2 Å from the initial structure. This RMSD value is significantly lower than is seen in the adsorption of the 9$^{\text{th}}$ domain alone, where the two replicas oscillate around 3 Å and 4 Å. Similarly, the 10$^{\text{th}}$ domain is just as stable on average - with RMSD value of 2 Å. It should be noted that the RMSD value was calculated with respect to the energy-minimised initial structure which is forcefield specific. However, this would not affect the overall conclusions. The domains are significantly more stable, and their RMSD values do not indicate that this would change.

The RMSD values do not indicate in any way that either of the domains will unfold, which is in large contrast to the simulations of FnIII$_{9\text{-}10}$ with CHARMM36 forcefield where the two domains did partly unfold (Figure 6.2). It is concluded that, the new forcefield which was created to improve the modelling capabilities of the disorder in proteins, actually decreases the disorder of this particular domain-pair.

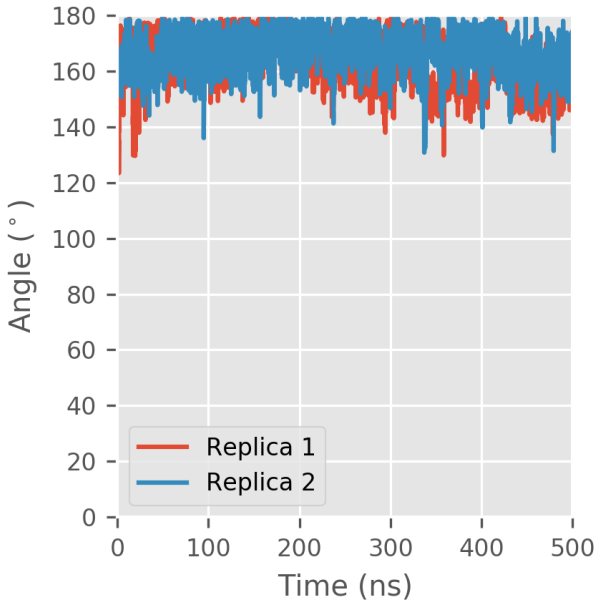


Figure 6.8: The RMSD of each domain measured separately in the FnIII$_{9\text{-}10}$ tandem and the udpated forcefield CHARMM36m. Both domains remain surprisingly stable throughout the simulation. The RMSD was measured with respect to the energy minimised structure with the same forcefield.

The angle between the two domains indicates that the domains interact less with each other at the temperature of 300K (Figure 6.8). In each of the two replicas the two domains remain in line, with the angle close to 180°. Therefore, the CHARMM36m forcefield more accurately describes the behaviour of the FnIII$_{9\text{-}10}$ [154, 155] and is better suited for the analysis of how the two domains can be arranged together.

The low RMSD values of each domain in FnIII$_{9\text{-}10}$ and the angle between the two domains with the CHARMM36m forcefield are most likely closely connected. The interface between the two domains affects their stability, as well as the ability of the domain to bend. It should be further noted that at the physiological temperature, the two measured values are likely to see an increased range and standard deviation.

6.5 Discussion

The tandem FnIII$_{9\text{-}10}$ simulations with the CHARMM36 forcefield led to the breakdown of the tertiary structure of the 9$^{\text{th}}$ domain. The partial unfolding was unexpected as the domain is

known to be stable in these conditions. For example, the heat capacity of the 9th domain (in the $FnIII_{8\text{-}9}$ complex) was found to be around 50°C [154]. In addition, NMR of the murine $FnIII_{9\text{-}10}$ complex indicates that the tertiary structure of each domain is stable [155]. Interestingly, the latter study shows that the two domains are not always aligned as the crystal structure suggests (PDB: 1FNF). Yet in the tandem $FnIII_{9\text{-}10}$ simulations, in both replicas, the domains have been observed to bend and interact with each other a lot. In one replica, the angle between the two domains was consistently around 80°, which means ample opportunity for cross-domain residue-residue contacts. This led to the hypothesis that the interactions between the 9th and the 10th domain is what leads to the partial unfolding.

To check whether the interactions between the two domains are responsible for the partial unfolding of the 9th domain, each domain was simulated alone, in the same conditions. The lone domains were found to be significantly more stable having lower RMSD values. The 9th domain also kept its tertiary structure. Therefore, it was concluded that it is interactions between the two domains in $FnIII_{9\text{-}10}$ simulations that leads to the loss of 9th domain's tertiary structure.

Another group performing molecular dynamics simulations of a bigger fragment, $FnIII_{8\text{-}10}$, might have run into a similar issue [93]. In that study, the authors used a structure extracted at 60 ns due to a "heavy bend between the ninth and tenth modules", even though the simulation in bulk was carried out for 100 ns.

CHARMM36m, the latest forcefield in the CHARMM family improves the treatment of the intrinsically disordered proteins (IDPs). The new forcefield was used to simulate the same complex $FnIII_{9\text{-}10}$. Not only was no loss of tertiary structure observed, but also each domain was found to be even more stable than when simulated alone with the CHARMM36. This shows that the new CHARMM36m is more suitable for the analysis of the interdomain $FnIII_{9/10}$ orientation, and for this reason CHARMM36m is used in the next chapter.

Chapter 7

Interdomain Orientation of $FnIII_{9\text{-}10}$ with CHARMM36m

In the last chapter, CHARMM36m was found to more accurately represent the interdomain relationship between the 9^{th} and 10^{th} domain. The previous version of the forcefield, CHARMM36, led to partial unfolding of the 9^{th} domain, which is refuted by experimental evidence. For this reason, in this chapter the CHARMM36m is used to analyse the relationship between the two domains. In Chapter 5, while adsorbing on the hydrophobic surface, the two domains changed their orientation with respect to each other. This suggests a possible conformation of the $FnIII_{9\text{-}10}$ which can adsorb with both domains to the EA SAM. Here, the interdomain relationship between the domains $FnIII_{9\text{-}10}$ is studied by creating 24 different configurations with each having the 10^{th} domain rotated around its principal axis.

7.1 Simulations & Methods

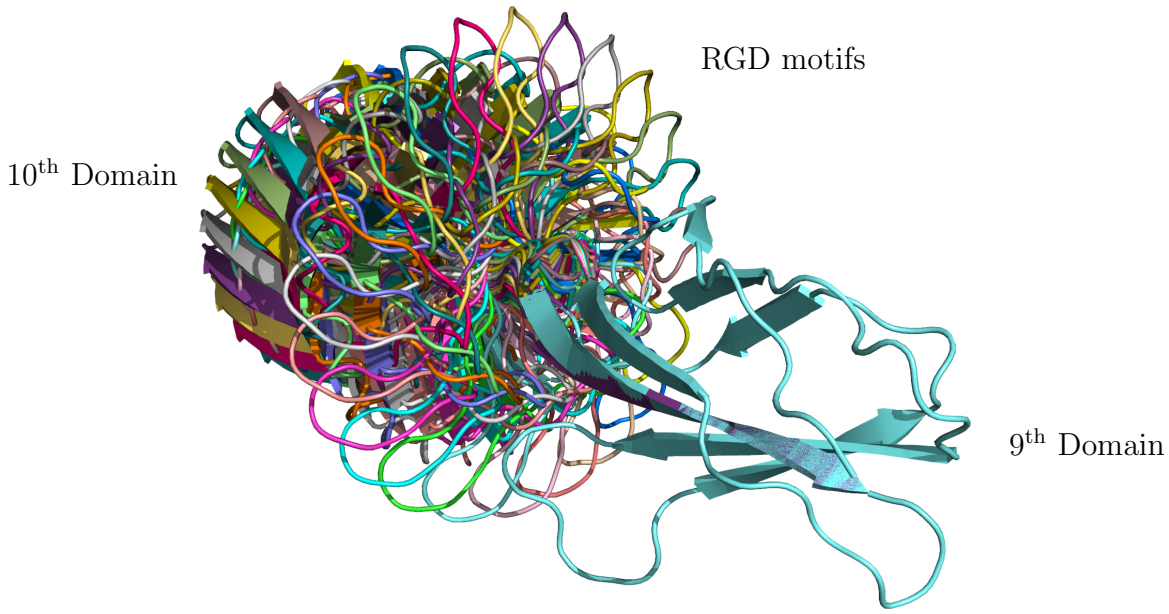


Figure 7.1: The initial 24 rotations of the $FnIII_{10}$ with respect to the $FnIII_9$, which were used to probe the interdomain preferences of $FnIII_{9\text{-}10}$

A set of 24 configurations of the $FnIII_{9\text{-}10}$ was created (Figure 7.1). For each consecutive system the 10th domain, which consists of residues 1416 - 1509 (inclusive), is rotated by an additional 15° around its principal axis. The last configuration is equivalent to the initial crystal structure and is therefore discarded. Each system was solvated and modelled using the CHARMM36m TIP3P water model [116]. The protein and the neutralising atoms were modelled with the CHARMM36m forcefield [120].

Each system was equilibrated via a series of simulations. First a steepest descent minimisation with a maximum force target of 500 kJ mol^{-1} nm^{-1} was performed. Then a series of simulations using the NVT ensemble with Nosé-Hoover thermostat were performed, where in each subsequent simulation was run at a higher temperature (such that simulations were performed at T = 100K, 200K, 250K and 310K). Each of these NVT simulations were run for 20 ps. The production simulation used the NPT ensemble with isotropic Parrinello-Rahman barostat applied every 5 ps. The production simulation used Nosé-Hoover thermostat at the physiological temperature 310K and each of the 24 systems was simulated for 100 ns. The other configuration details are given in Chapter 3.

The analysis of a twist between the two domains $Fn_{9\text{-}10}$ was carried out using super-dihedrals defined in Chapter 5. In order to compare the final states of each system, the last 5 ns (with the step 100 ps) were transformed into distance maps. Each distance map contained the minimum distances from the heavy atoms of each residue to the heavy atoms of every other residues. The distance maps were collated for each trajectory into a single distance map by averaging the distances over time. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [156] was used to extract the clusters from the distance maps of the 24 systems. As the measure of similarity between the distance maps, root mean square (RMS) difference between the corresponding residue-residue distances was used. In order to distinguish between the structural RMSD and the comparison of distance maps, the "RMS difference" will be used in the context of distance map comparison. The maximum RMS distance between two distance maps to be considered in the same cluster was set to 1.5 Å. Different values were tried and clusters generated with them were visualised in order to find the most meaningful structural clusters.

The room mean square deviation (RMSD) was calculated between the structure of a domain and its equilibrated version which was used at the beginning of the production simulation.

7.2 Stability

The stability of the 9^{th} and the 10^{th} domain is assessed independently in each system using the root mean square deviation (RMSD) of their structures over time (Figure 7.2). The figure is divided into two halves: the RMSD of the 9^{th} domain is presented on the plots with the purple background, and for the 10^{th} domain, the khaki-coloured background is used.

The 9^{th} domain is perfectly conserved across the different rotations of $FnIII_{9\text{-}10}$. Across the 24 simulations, the RMSD value never exceeds 3 Å from the energy-minimised crystal structure (PDB:1FNF [1]). For the 10^{th} domain occasional spikes in the RMSD are seen, particularly in the r285 system. However, on average, the structure is highly stable across the different $FnIII_{9\text{-}10}$ rotations. Furthermore, a consecutive series of rotated systems remains highly stable.

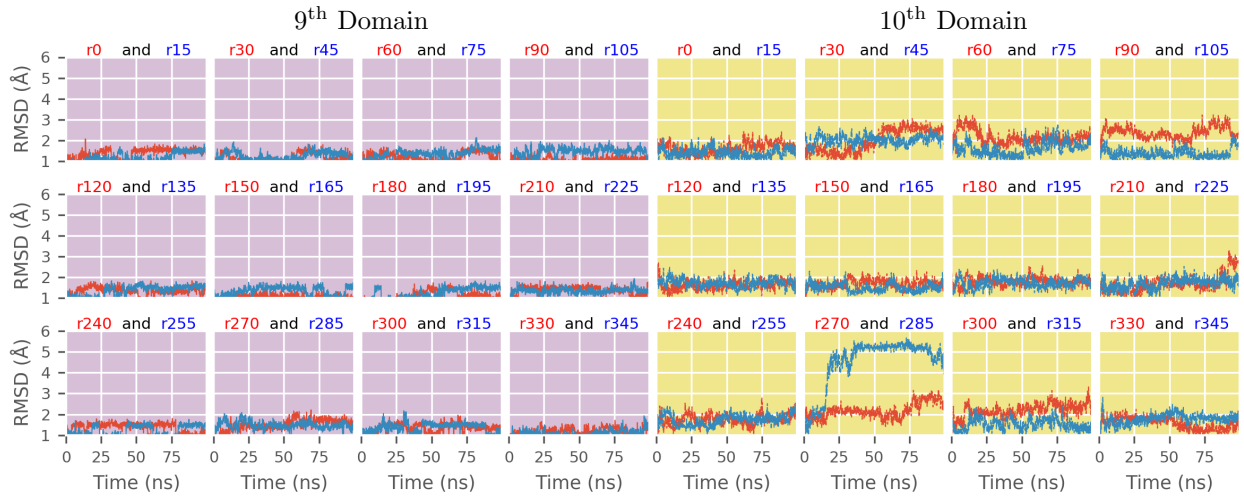


Figure 7.2: The RMSDs of the structure of the 9th domain (purple background) and of the 10th domain (khaki background). Each plot presents two systems with a different initial rotation of the 10th domain. For example, the first plot represents the system with a rotation 0° system (r0) with the red title matching the red curve. The system r15 (15°) is coloured blue which matches the colour of the curve.

For example, the systems r315 - r15 rarely reach an RMSD value of 3 Å. Similar high stability is observed in the peptides which have an initial rotation ranging from 120° to 255°.

There is also one series of consecutively rotated systems that exhibit a small increase in the RMSD value: the systems r300 - r105. However, the RMSD values are still very low and only occasionally reach 3 Å. There is only one simulation with a larger RMSD, r285, which fluctuates around 5.5 Å. Yet, a visual inspection of the 10th domain shows that it does not lose its secondary structures. Setting r285 aside, the RMSD of the 10th domain is, on average, only marginally larger than that of the 9th domain. This is likely due to the longer RGD loop F/G which is free to move (see Figure 1.2).

There appears to be a weak relationship between the RMSD of the 10th domain and the rotation of the domain. Overall, however, both domains are highly stable, and this allows for measurement of the angle between the two domains, and of the rotation of one domain with respect to the other. These two are carried out in the following sections.

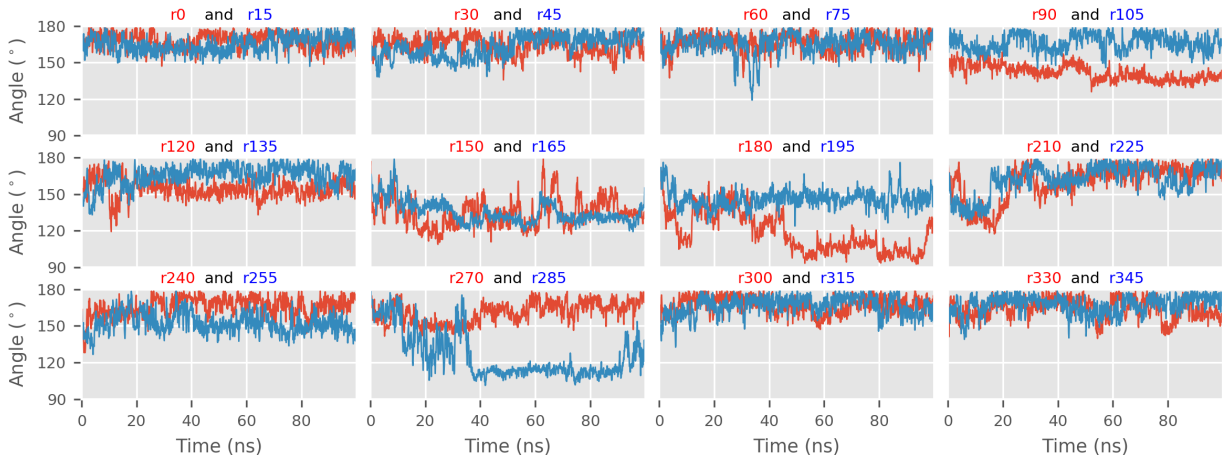


Figure 7.3: The interdomain angle between the centres-of-mass of the 9$^{\text{th}}$ domain, the linker and the 10$^{\text{th}}$ domain. The title number following the 'r' letter indicates the degree by which the 10$^{\text{th}}$ domain was rotated initially, whereas its colour matches the line plotted.

7.3 Angle

In this section the angle between the 9$^{\text{th}}$ domain, the linker and the 10$^{\text{th}}$ domain is discussed (Figure 7.3). The variation of an angle between the domains depends on the initial rotational configuration of the FnIII$_{9\text{-}10}$ complex. Firstly, one might distinguish the consecutive series of simulations r300-r75 which consistently fluctuate around the angle of 165°. The first system that diverges significantly from this is r90, which ends with the angle of 140°.

In r150 and r165 the angle between the two domains fluctuates mostly around 135° during most of the simulations. The next rotated system, r180, shows even more variation - during the second half of the simulation the angle is between 90° and 120°, ending at 120°. The system r255 fluctuates on average around 150°. The last system, r285, being relatively close to a right angle is similar to r180. In that system, the angle between the two domains first drops to around 100° and towards the end if returns to the angle 120°. The interdomain angle is often close to 90° in the systems r180 and r285 which provides ample of opportunity for domain-domain contacts. Despite this, the domains conserved their tertiary structures. By comparing Figures 7.2 and 7.3, there is no obvious correlation between the interdomain angle and the RMSD of the domains.

In the majority of the 24 systems, the interdomain angle remains close to the straight angle

fluctuating around 165°.

7.4 Rotation

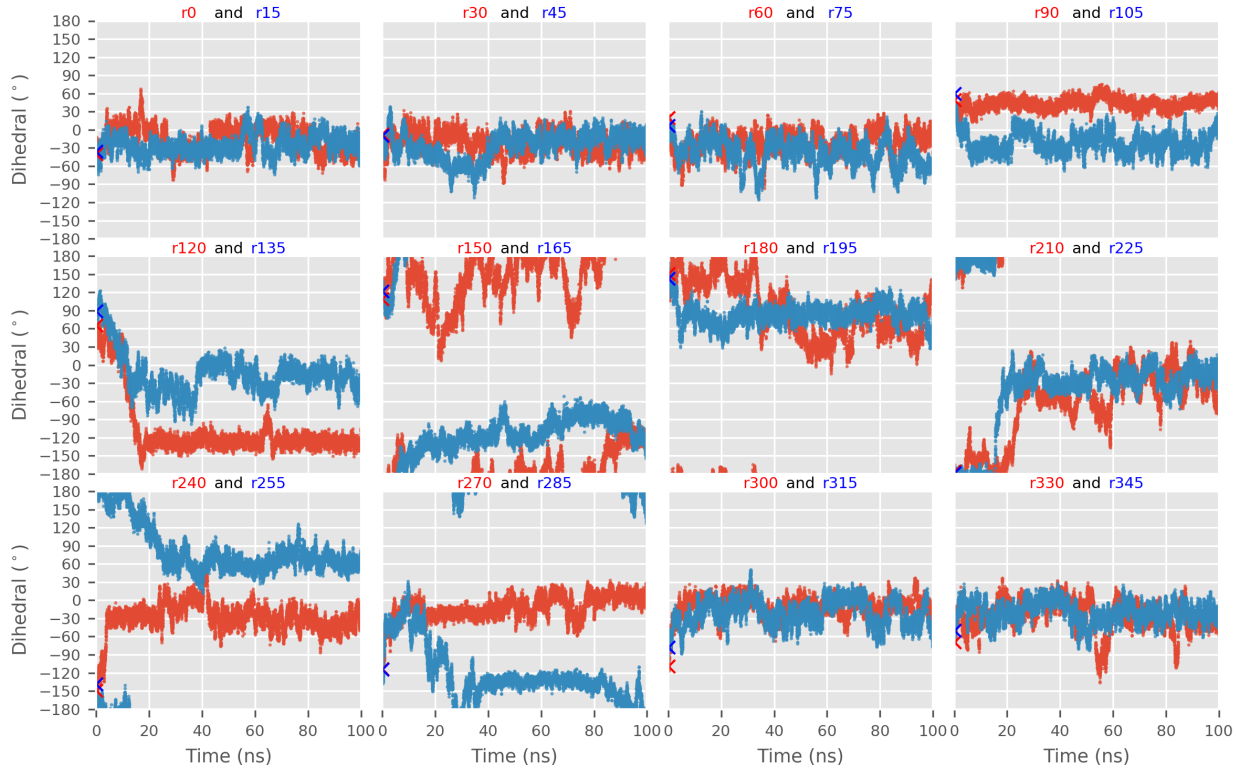


Figure 7.4: The interdomain rotation described with a super-dihedral made up of the centres-of-mass of four residues (1396-1345-1434-1486), which are placed on the four different β-sheets in the two $FnIII_{9\text{-}10}$ domains (see Chapter 2 for definition). The number in the title following the r letter indicates the initial degree by which the 10^{th} domain was rotated, whereas the colour corresponds to the plotted line. The initial starting point for each system is marked with a symbol X.

The interdomain rotation was quantified with a super-dihedral that relied on the centres-of-mass of four residues: Ser1396-Val1345-Leu1434-Thr1486. These residues represent the centres of each β-sheet of the 9^{th} and 10^{th} domain, respectively. For each of the 24 systems the super-dihedral angle as a function of time is shown in Figure 7.4. Please note that the super-dihedral does not perfectly correspond to the initially introduced rotation of the 10^{th} domain.

The system r0 starts close to the initial interdomain orientation in the crystal structure PDB:1FNF [1]. In this case, the two domains remain in a similar relative position for most of the simulation. A similar pattern is found in more systems: once again the consecutive series of rotated

systems r300 to r75 behave the same way, in this case fluctuating around the dihedral of -30°. These systems with the 10th domain rotated by 300° to 75° were also found to have a stable interdomain angle, meaning that the most stable interdomain orientation is the one found in the crystal structure.

The first system that departs from the -30° minima is r90. In that system, the dihedral stays consistently between 30° and 60°. This is while the interdomain angle drops to around 135°. There is another system that behaves very similarly: r255. In that system both the dihedral and the interdomain angles are similar. Furthermore, r180 and r195 follow a similar trends. Out of these two, r195 transitions to 30-60° right at the end of the simulation. Once again, this r195 also has an interdomain angle that often is below 150°. The r180 dihedral fluctuates largely between 30° and 120° and towards the end if found closer to the latter. However, its interdomain angle is very different. Thus, the three systems are similar to each other.

Another system worth noting for its dihedral fluctuating steadily around the same value is r120, which is between -120° and -150°. To some extent, two other systems, r150 and r165, end with the same interdomain rotation. These two systems have a smaller interdomain angle, often around 135°, which is at odds with the r120 interdomain angle which fluctuates around 150°.

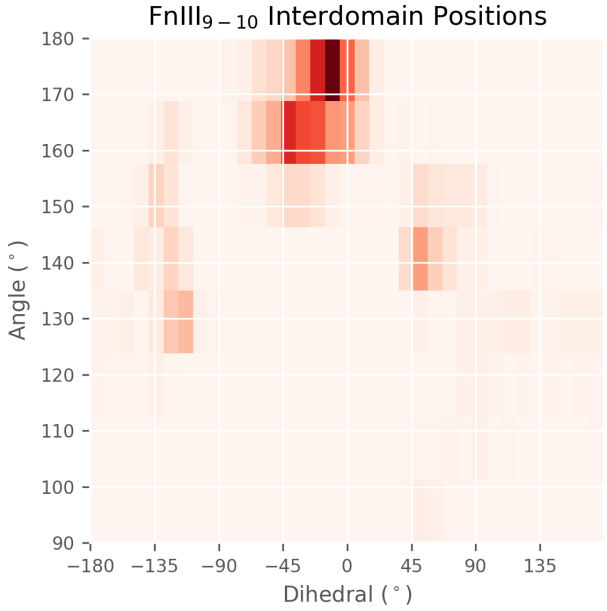


Figure 7.5: The heatmap visualises the interdomain angle and the dihedral angle found at the ends of the simulations. It combines the last 5 ns of each of the 24 systems.

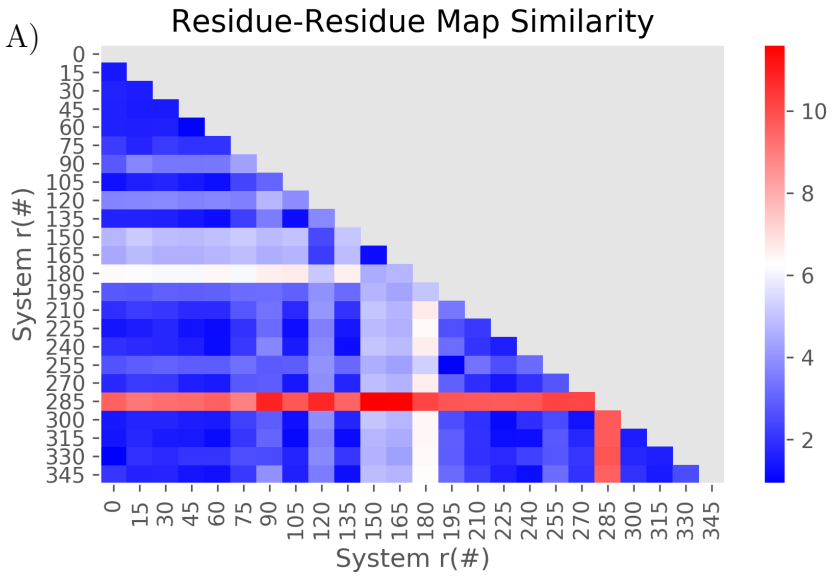
For the last 5 ns of each of the simulations, the dihedral angles were paired with the interdomain angles and converted into a heatmap shown in figure 7.5. The heatmap is dominated by a single island centred around the -25° dihedral and the interdomain angle of 170°. In addition, there are two different areas that are separated from the main island. The first one is at around 50° dihedral and 140° angle. It has been previously suggested that this is due to systems r90, r180, r195 and r255. The second area is at the -125° dihedral and 130° angle, which was previously described by r120, r150 and r165. Furthermore, it appears that the dihedral angle can be used to uniquely distinguish the different conformations.

It is not clear, however, whether the islands represent steady or transient states. The same rotations of the 10^{th} domain according to dihedral might mean in practice very different residue-residue interactions depending on the interdomain angle. Moreover, the timescale of the simulations is limited. In one case, r150, which undergoes significant changes before finishing at around the dihedral of -120°, extending the simulation could affect the final state. In the next section, in order to better understand the similarities and differences between the simulations, and what stabilises the differently rotated systems, residue-residue interactions are used to understand the discussed interdomain angle and super-dihedral patterns.

7.5 Clusters

The analysis of the dihedral angles shows that $\text{FnIII}_{9\text{-}10}$ assumes three different rotational conformations. These conformations should mean that the two domains interact with each other in some specific ways to ensure that they stay in the observed states. This implies that residue-residue interactions can be used to find which systems are similar and as well as which interactions make them so. Here, the similarities between the different conformations of $\text{FnIII}_{9\text{-}10}$ are investigated using distance maps. These maps contain minimum distances from the heavy atoms of each residue to the heavy atoms of each other residues. Each distance map was created by averaging the residue-residue distances in the last 5 ns of each simulation. Then, the RMS difference was used to compare these averaged distance maps. The results are presented in Figure 7.6A.

A)



B)

Clusters
0-75, 105, 135 210-240, 270, 300-345
120, 150, 165
195, 255
90
180
285

Figure 7.6: **A)** The RMS difference between the contact maps of each rotation system. A lower value means that the contact maps are similar. The description of how two contact maps are compared to each other is provided in the methods section. **B)** The clusters computed with DBSCAN applied on the comparison matrix. The preceding letter 'r' is omitted.

There are two systems that stand out by being distinctly different to all other systems. One of them, r285, is different from every other system and stands out in the red colour. In this system the 10$^{\text{th}}$ domain has an usually large RMSD (Figure 7.2). Another system that is significantly different to most of the other systems is r180, with a RMS difference value of around 6 Å. However, r180 retains some similarity to a few systems: r120, r150, r165, r195 and r255. The first three, r120, r150 and r165, stand out on their own: r120 is more distant from all simulations except for r150 and r165. Moreover, these two systems are very similar to each other. Thus, it appears that the three r120, r150 and r165 form a hub. The latter systems, r195 and r255, are also close to each other. Therefore, r180 is more similar to these two hubs than to any other system.

In order to confirm the existence of these hubs clustering with DBSCAN is carried out: the results are shown in Figure 7.6B. The outcome is 6 clusters with the first three including more than one member. The first cluster counts 16 simulations and represents the crystal structure FnIII$_{9\text{-}10}$ orientation (See Figure 7.6B). The second cluster contains r120, r150 and r165, whereas the third cluster shows a the pair r195 and r255. The systems in the clusters were also previously shown to have similar super-dihedral and interdomain angles. The last

three clusters contain one system each, and they are r90, r180 and r285.

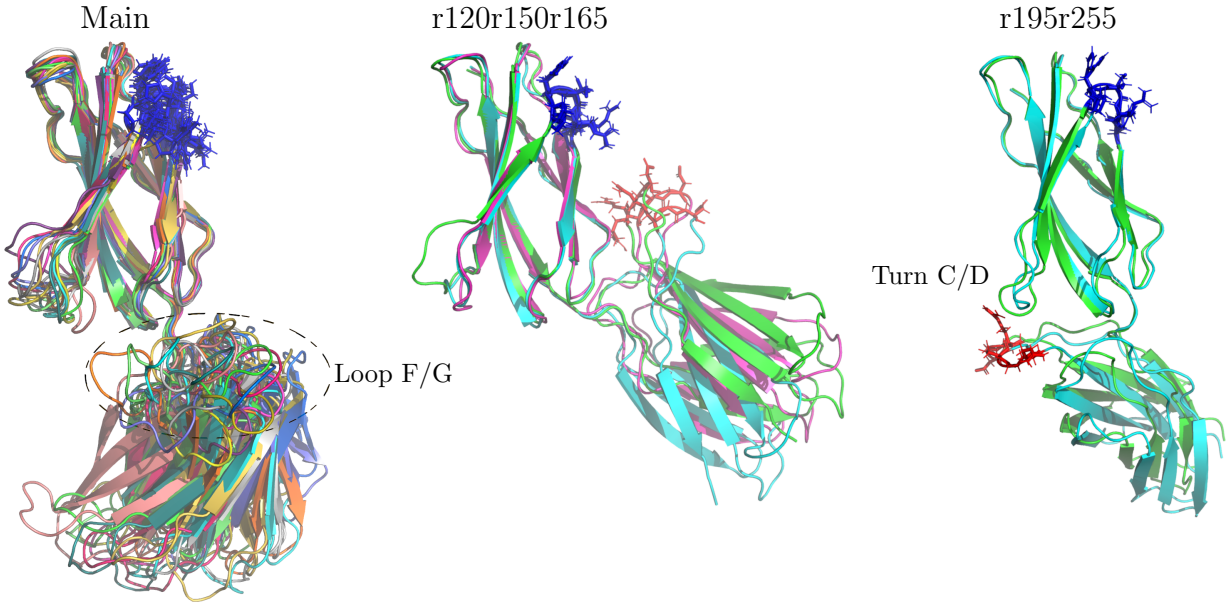


Figure 7.7: The three clusters counting more than one member visualised using the last frame for each system. In each, the structures are superimposed using the 9th domain (Cα - atoms). Each structure is coloured differently. The 9th domains in each cluster (top domain) are oriented the same way to highlight the different positions of the 10th domain. Red licorice represents RGD motif and is omitted in the main cluster for clarity. The PHSRN motif is visualised with blue licorice.

Let us look at each cluster by taking a final structure from each and superimposing it using the 9th domain. This way the orientation of the 10th domain is highlighted. The three clusters with more than one member are shown in Figure 7.7. Each cluster represents a different orientation, but within each of the clusters the $FnIII_{9\text{-}10}$ conformations are almost the same. The main cluster, which counts 16 members, shows that the 10th domain has a bit of room to wiggle. This was previously shown by the range of dihedrals in Figure 7.5.

In the second cluster, called r120r150r165, the characteristic element is the RGD motif (red licorice) and its proximity to the PHSRN motif (blue licorice). The dihedral angle of the last frame in the r120 system is -120°, which was previously identified as one of the conformations in the dihedral angle heatmap (Figure 7.5).

In the case of the r195r255 cluster, the RGD loop F/G is close to a different β turn (for notation Figure 1.2). It is more similar to the main cluster with the biggest difference being that the RGD loop F/G is facing in the opposite direction, and appears to interact with the turn C/D

in the 9th domain. Whereas this interaction is missing from the main cluster, the RGD loop in that cluster is sometimes oriented in the same direction. The dihedral angle in the final structure in r195 is 88.50°. However, this dihedral angle does not clearly indicate that there is a minima (Figure 7.5), with the closest minima being at 50°.

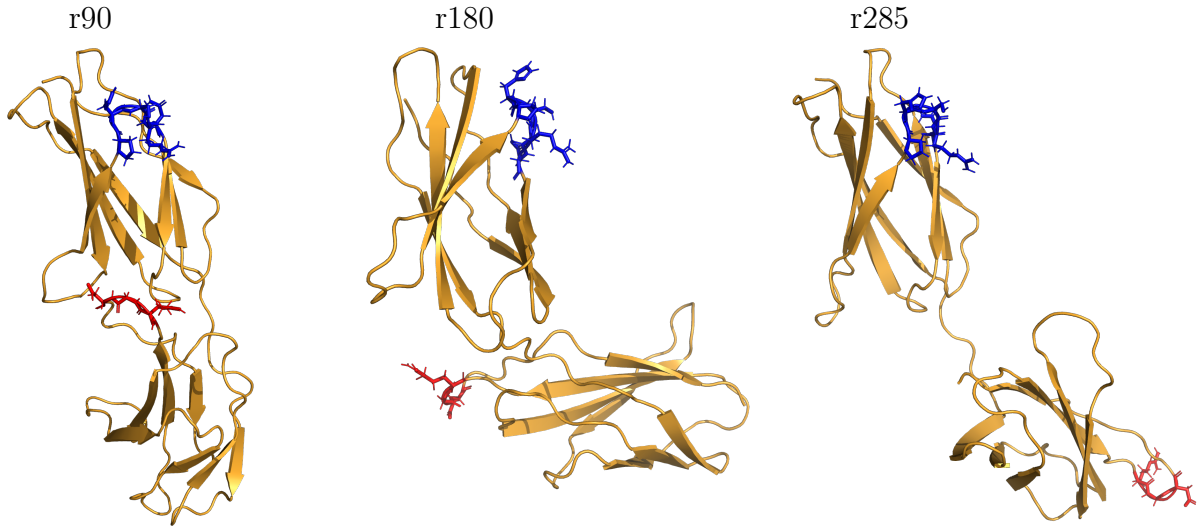


Figure 7.8: The three single-member clusters visualised using the last frame for each system. The 9th domains (top domains) in each cluster are oriented in a similar way to highlight the different positions of the 10th domain. The red and blue licorice represent the RGD and PHSRN motifs, respectively.

In addition to the three group-clusters there are also the three single-member clusters shown in Figure 7.8. For clarity, the 9th domain was reoriented to highlight the orientation of the 10th domain. The first cluster, r90, has the RGD loop F/G oriented towards the turn C/D in the 9th domain. These two loops are also close in the r195r255 cluster. The final dihedral angle in system r90 is 50° which puts it closest to r195r255. With the loop and the dihedral being in between the main and r195r255 clusters, the r90 system is likely an intermediate between them. The similarity matrix in Figure 7.6 further adds weight to this idea by showing that r90 is as similar to r195r255 as it is to many other systems in the main cluster.

The system r180 also appears to be an intermediate state between the main and r195r255 clusters - the same two loops are only slightly farther away from each other. The main difference is the rotation of the 10th domain. With a dihedral angle 122° the system is still closest to the cluster r195r255.

The last cluster is the outlier r285 - and the visualised structure explains this. The linker between the two domains is extended affecting many of the global residue-residue distances. This lack of a buried interface between the 9^{th} and 10^{th} domains also explains why the angle between the two domain has changed so much. Whereas the secondary structures are conserved in both domains, the linker significantly affects the RMSD of the 10^{th} domain.

Two out of the three single-member clusters appear to be intermediate states, whereas in the third cluster (r285) the domain-domain interface is lost due to an elongated linker. Therefore in the next section I focus on the main cluster and its comparison to the r120r150r165 and r195r255 clusters.

7.6 Residue-Residue Interactions

To understand the difference between two clusters, each cluster was collated into a single distance map by taking the minimum residue-residue distance across the cluster. This way residue-residue interactions are overrepresented in each distance map. Next, the distance maps are reduced to a contact map by accepting only distances below 6 Å. The distance captures most of the van der Walls interactions for the heavy-heavy atoms distances. Then the contact maps are compared by calculating the difference between them. This way the unique residue-residue contacts that distinguish the two clusters are highlighted.

7.6.1 Main and r120r150r165 Clusters

The cluster r120r150r165 is compared to the main cluster in Figure 7.9. Here, the focus is on the bottom-left quarter of the graph which represents the interactions between the two domains and is separated with the grey dashed line. It shows that some residue-residue contacts between the two domains are present only in the main cluster (red), and others only in the r120r150r165 cluster (blue). Furthermore, for each of the clusters the residues involved in the interactions are visualised as licorice in the complementary Figure 7.10.

One might spot first the blue patches in a rectangle (dashed-line) labelled C1 on the x-axis

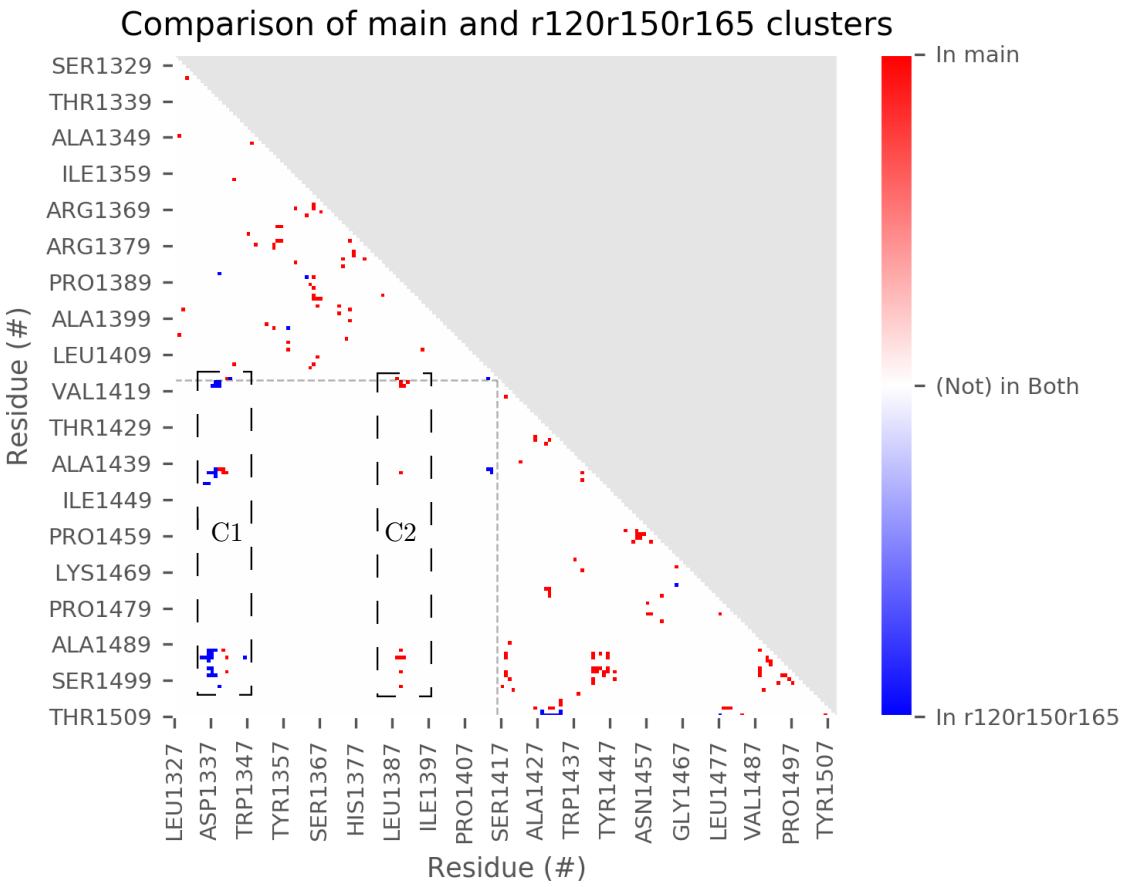


Figure 7.9: The difference between the contact map of the main and r120r150r165 clusters. The minimum value was used to collate contact maps in each cluster. The red areas represent residue-residue contacts that are present in the main cluster but are absent from the r120r150r165 cluster, whereas the blue colour indicates the reverse. White colour indicates either the existence or lack of contacts in both clusters. The bottom-left square separated by a grey dashed line describes the interdomain differences.

centred around Asp1337 involving residues 1334-1339, which reside on the β-sheet$_{\mathrm{A}}$ and turn A/B in the 9$^{\mathrm{th}}$ domain. This site contains the majority of new interactions with the 10$^{\mathrm{th}}$ domain in the r120r150r165 cluster. It involves three different regions on the 10$^{\mathrm{th}}$ domain: around Asp1418 (linker area, green), Val1442 (turn$_{\mathrm{B/C}}$, yellow), and Gly1494 (RGD loop F/G, red colour). The latter loop uses the RGD residue Arg1493, with its guanidine group being close to Asp1334 (4.9Å), Ser1336 (3.8 Å) and Asp1337. The two negatively charged aspartic acid residues appear to stabilise this position of the 10$^{\mathrm{th}}$ domain and form the r120r150r165 cluster. The visualised final structure of r120r150r165 cluster shows that this is the main hub of contacts in this cluster. These new contacts on the 9$^{\mathrm{th}}$ domain displace several contacts as shown by the close red spots in the rectangle C1 - these engaged the residues in between the

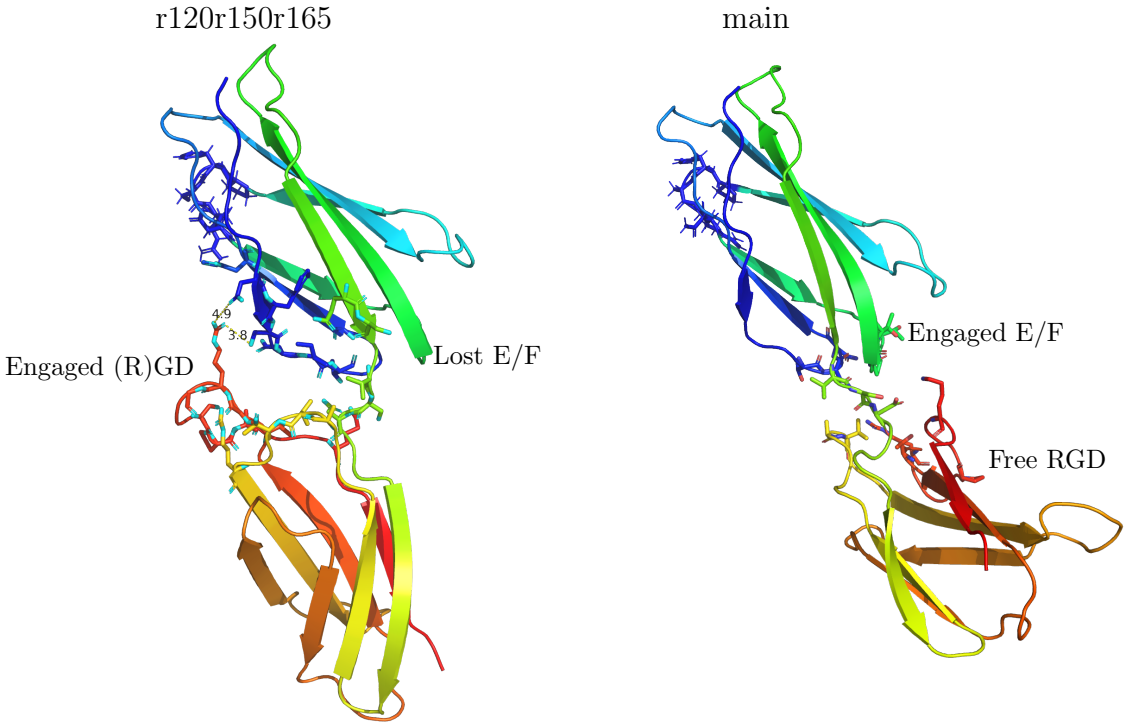


Figure 7.10: Comparing the final states of the r120r150r165 and main clusters as represented by r120 and r0, respectively. On the left (r120r150r165) licorice is used to show residues involved in the interactions across the two domains which are present in r120r150r165 but not the main cluster (Figure 7.9, blue). On the right, it is the other way around (Figure 7.9, red). The PHSRN motif is represented with blue licorice in the top domain and the polypeptide is coloured with a blue-red rainbow along the chain.

two domains and are also present in the crystal structure. As the 10^{th} domain bends to the side, they are lost.

Then, the contact map shows that very close to the linker on the 9^{th} domain (Gln1413 and Ser1414) there are a few new contacts in the cluster r120r150r165, one in the linker area and one with the $\text{turn}_{B/C}$ with residues Ala1441 and Val1442.

From the point of view of the main cluster, several residues around Pro1389 on the 9^{th} domain stop interacting with three different locations on the 10^{th} domain, as shown by the three red regions in Figure 7.9 in the rectangle C2. The residues Thr1388, Pro1389, Gly1390 and Thr1391 reside on the turn from β - strand E/F, and are absent from the cluster r120r150r165. Furthermore, in the main cluster, the RGD motif is not engaged in any interactions with the 9^{th} domain (Figure 7.10A).

7.6.2 Main and r195r255 Clusters

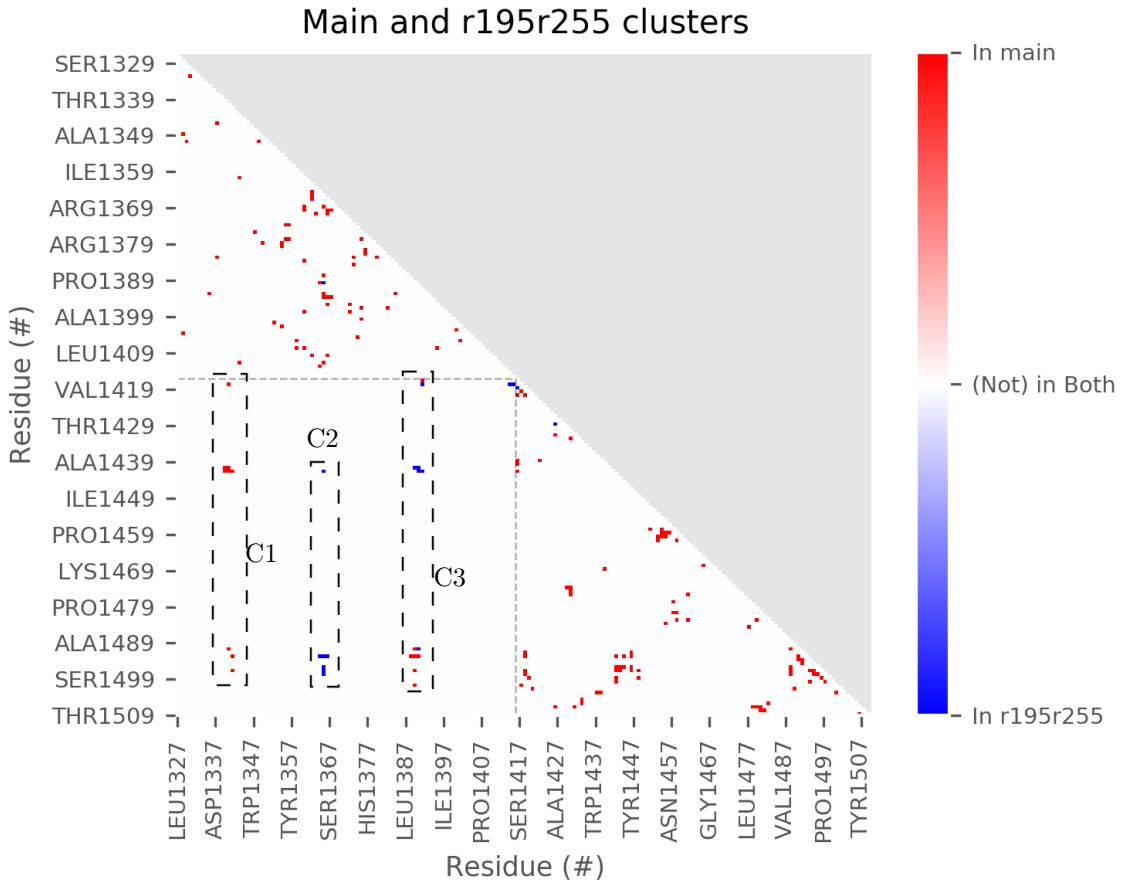


Figure 7.11: The difference between the contact map of the main and r195r255 clusters. The minimum value was used to collate contact maps in each cluster. The red areas represent residue-residue contacts that are present in the main cluster but are absent from the r195r255 cluster, whereas the blue colour indicates the reverse. White colour indicates either the existence or lack of contacts in both clusters. The bottom-left square separated by a grey dashed line describes the interdomain differences.

The r195r255 cluster is compared to the main cluster, with the difference in the contact maps shown in Figure 7.11 and the complementary snapshots of the two representative structures of the two clusters in Figure 7.12.

In comparison to the main cluster, in r195r255, the residues Thr1339, Ala1340 and Asn1341 (area C1) on the turn A/B in the 9^{th} domain lose contact with three different regions in the 10^{th} domain: around Arg1493, Val1442 and Asp1418. The three residues on the 10^{th} domain reside on different turns/loops: green linker, orange RGD loop, and yellow loop. These three turns/loops are still engaged in r195r255 but with different residues.

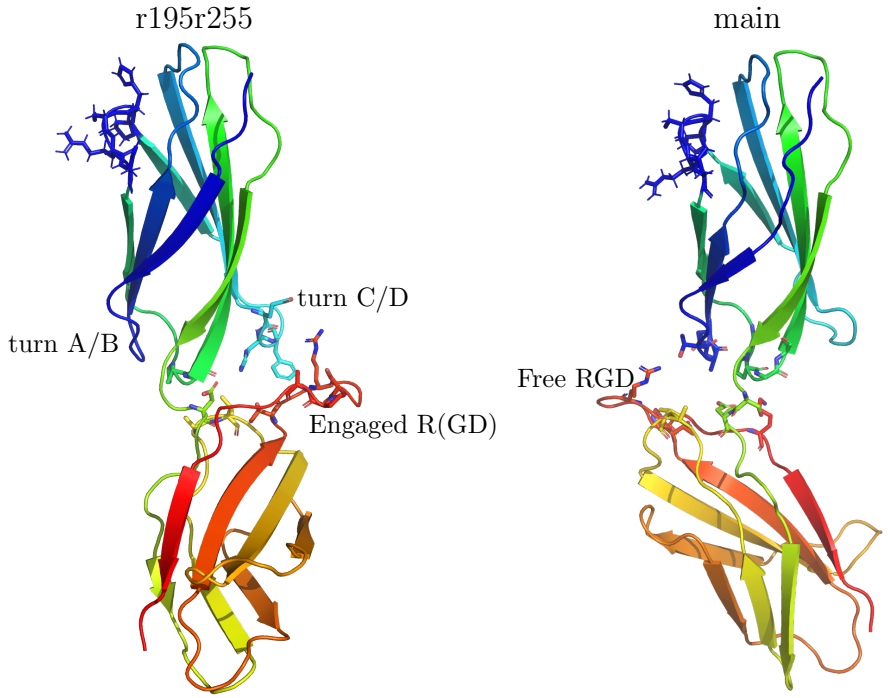


Figure 7.12: Comparing the final states of the r195r255 and main clusters, which are represented by r120 and r0, respectively. On the left, licorice is used to represent residues interacting across 9th and 10th domains in the r195r255, but not in the main cluster (Figure 7.11, blue). On the right, it is the other way around (Figure 7.11, red). The PHSRN motif is represented with blue licorice in the top domain and the polypeptide is coloured with a blue-red rainbow along the chain.

The area C2 in the contact map highlights residue-residue contacts which are present in r195r255 but not the main cluster. These contact involve residues Glu1364, His1365 and Phe1366 in the 9th domain which reside on a turn C/D (Figure 7.12, cyan colour). They are found close to two different regions on the 10th domain: around Arg1493 and to a lesser degree Val1442 (visualised as licorice). One of the new contacts in the interdomain area is Arg1493 with Glu1364 which, due to the opposite charges, likely contributes to the stability of this cluster.

In comparison to the main cluster, in the area labelled C3 three residues, Thr1388, Pro1389 and Gly1390 on the 9th domain on the turn E/F, move away from Arg1493 and Pro1497. Instead, in r195r255, they are found close to Ala1441 and Val1442 which are on the turn$_{B/C}$ (yellow).

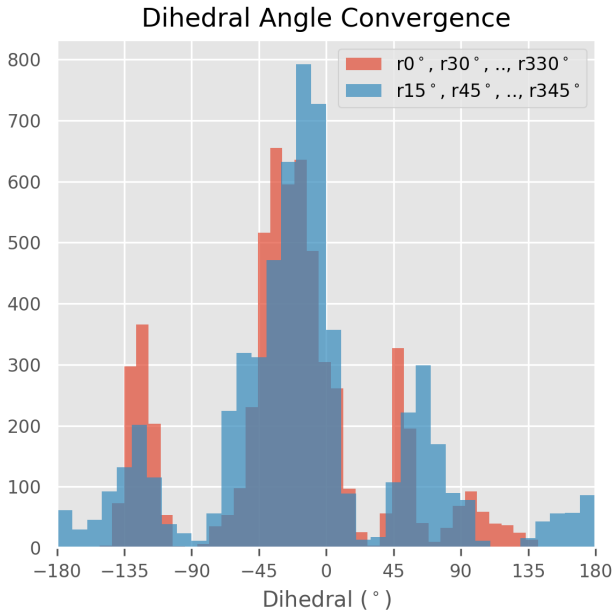


Figure 7.13: The convergence of the dihedral angles. The 24 systems are split into two subsets, as illustrated by the legend. The last 5 ns of each simulation was used. -15°, -30° correspond to r345°, r330°, etc.

7.7 Rotation Convergence

In this section the convergence of the dihedral angles is discussed. The 24 systems were divided into two subsets of 12 simulations: systems r0°, r30°, .. , r330°, and systems r15°, r45°, ..., r345°. The histogram of each subset was plotted in Figure 7.13.

The major peak in both subsets is centred around -20° which represents the main cluster with the rotation at which most of the simulations finish. These major peaks range from -60° to 15° which contains the dihedral orientation found in the crystal structure, which is 0° (PDB:1FNF). However, the main peak in the subset r15°, ..., r345° consists of two peaks: a smaller one centred at the -55° dihedral and the big one centred around the -15° dihedral.

There are two other smaller peaks in each subset that clearly overlap: one at -125° and one around the 55° dihedral. The former contains the r120r150r165 cluster which explains the higher peak in the subset r0°, ..., r330°.

The second pair is at around 50° and 65° for the two subsets. The latter refers to the 88° dihedral angle found in the r195r255 cluster. Removal of this cluster leads to the disappearance of the 65° peak in the visualised subset r15°, ..., r345°. The peak at 50° represents only the r90 system,

further adding evidence to the idea that it is similar to the r195r255 cluster. The same trend is followed by the flatter peak at 95° which represents r180.

After splitting the simulations into two subsets, the overall patterns in each subset are very similar. It appears that a few simulations, including r90 and r180, have not fully converged. However, the qualitative discussion shows that the major patterns are reproduced.

7.8 Discussion

In this chapter the interdomain orientation of the $FnIII_{9\text{-}10}$ was analysed with the use of the latest forcefield CHARMM36m. Together, 24 different interdomain orientations were simulated for 100 ns to find the preferred conformations. The tertiary and secondary structures of each domain remained stable throughout all simulations, and the analysis of the interdomain angle and the twist showed that there is a major interdomain orientation preference. This preferred major orientation is that of the crystal structure (PDB:1FNF). However, there were exceptions: several simulations ended with different interdomain and super-dihedral angles.

The interdomain and super-dihedral angles analysis was further refined by the clustering algorithm DBSCAN for which residue-residue distances were used. Six clusters emerged out of which three were discarded: two as intermediate states, and one due to the loss of the interdomain buried interface. The three clusters counted consecutively 16, 3 and 2 members, which highlights the size of the main cluster. In the case of the r120r150r165 cluster, the RGD motif is directly involved in the electrostatic interactions with two aspartic acid residues while engaging a part of a β - strand on the 9^{th} domain. In the r195r255 cluster, on the other hand, the RGD motif is on the opposite side close to turn C/D. In contrast, in the main cluster, these areas are not involved in cross-domain residue-residue interactions.

The orientation in the cluster r120r150r165 is similar to the interdomain orientation of $FnIII_{9\text{-}10}$ that was first observed during the adsorption to the methyl SAMs in Chapter 5. Recall that in that chapter the interdomain angle dropped to around 130° in both replicas, and the dihedral angle is -40° and -140° for the methyl18 and methyl10 replicas, respectively. The latter

-140° dihedral with a 130° angle is very close to the cluster r120r150r165, showing that conformation can be reached during the adsorption to the hydrophobic surface. However, the protein fragment did not change its original interdomain orientation during the adsorption on the EA SAMs in Chapter 3.

In Chapter 6 I showed how the tertiary structure of $FnIII_{9\text{-}10}$ breaks down with the CHARMM36 forcefield which justified the transition to the CHARMM36m. The new forcefield significantly affects the interdomain orientation of the complex in bulk simulations, and further analysis shows that three different interdomain orientations can be assumed. Interestingly, one of them was seen previously in Chapter 5 during the adsorption of the fibronectin fragment to the hydrophobic surface.

Whereas long classical MD simulations can be used to probe interdomain orientation preferences [157], they are limited in their sampling and are computationally costly. Therefore, it is important to consider other approaches taken to characterise the interdomain orientation. An interesting approach is that of metadynamics [158], where the collective variable could be a distance map of the residues at the interface between the two domains. Another enhanced sampling method is that of simulated annealing. A different way of improving the sampling is with the use of coarse-graining [159], although that creates new challenges related to the parametrisation of the coarse-grained models [160].

In this chapter, the clustering was carried out using the residue-residue distances with RMS difference. Another approach worth considering is that of clustering high-dimensional descriptors. In this case, the descriptors could refer to metrics such as the previously defined domain-domain angle (bend) or dihedral angle (rotation). However, other metrics could be included, such as domain-domain distance, the number of hydrogen bonds or the interaction energy between the two domains.

With knowledge of the different interdomain orientations, and the degree to which CHARMM36m affects the $FnIII_{9\text{-}10}$ fragment, in following chapter I investigate how it affects the adsorption to the three substrates EA, MA and methyl SAMs.

Chapter 8

Adsorption of $FnIII_{9\text{-}10}$ to EA, MA and Methyl SAMs with CHARMM36m

In the previous chapters I showed that the forcefield CHARMM36m improves the representation of the tertiary structure stability of the $FnIII_{9\text{-}10}$ complex in bulk simulations. Whereas the effect of the forcefield has clear implications on the two domains in bulk water, its effect on the adsorption is less clear. In this chapter a new set of simulations are designed to find out how much of a difference the forcefield affects the adsorption on the three previously-studied surfaces: EA, MA and methyl SAMs.

	mMA	MA	mEA	EA	mMethyl	Methyl
Tilt (°)	64.74±0.35	61.4±0.26	62.00±0.71	61.6±0.25	61.38±0.89	57.5±0.77
Roughness (Å)	0.32	0.37	0.40	0.35	0.30	0.43

Table 8.1: The tilt (and its standard deviation) and the roughness of the surface described with the average standard deviation for the model surfaces (MA, EA and Methyl) and for the updated surfaces (mMA, mEA and mMethyl). The carbon atom preceding the functional group was used to measure the roughness. The calculations were performed on the frames during the last 10 ns of a simulation.

8.1 Simulations & Methods

For each of the surfaces, EA, MA and methyl SAMs, a system containing the SAM on top of a gold slab modelled with GOLP[141] was assembled as described in Chapter 3. These systems were used to obtain equilibrated states of the SAMs for the calculation of the tilt and the roughness of the surface (Figure 8.1). The systems were simulated without any protein and ions and the protocol used for these simulations is described in Chapter 3.

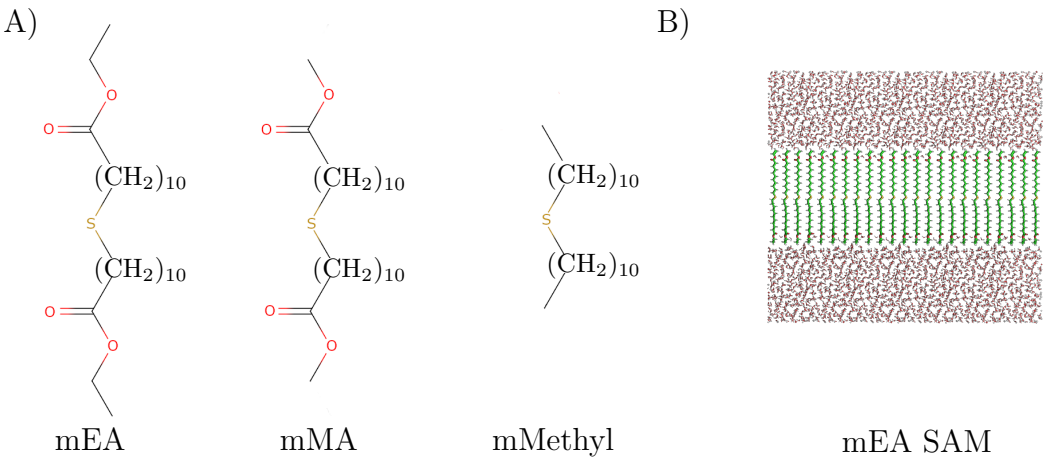


Figure 8.1: **A)** The new chains from which the new surface is assembled. They consist of the central sulphurs with the same carbon chain and functional groups at the top and at the bottom. To the right **B)** an assembled mEA SAM solvated in water.

In order to improve the performance and thus sampling, a new simplified system was assembled which excluded the gold slab due to its minor role in the adsorption on the SAMs. This relative insignificance of the gold slab was shown in the previous chapters where the fibronectin fragments did not even penetrate the self-assembled monolayers. The only function that the slab had was helping construct a realistic self-assembled monolayers surface. Here, after the removal of the gold slab, the tilt and the roughness of the surface were approximated in the

new systems. The chains are defined as S[$(CH_2)_{10}$R]$(CH_2)_{10}$R, where R is $-COOCH_3$ for EA, $-COOCH_2CH_3$ for the MA and $-CH_3$ for methyl terminated SAMs. The new chains are are called mEA, mMA and mMethyl SAMs, as shown in Figure 8.1A. These new chains allow us to use the PBC in all three dimensions.

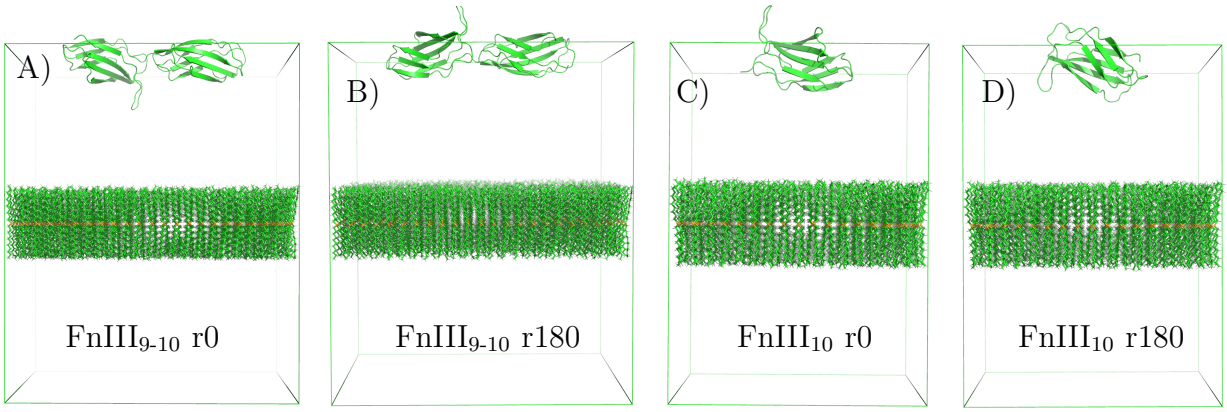


Figure 8.2: Visualised here are four out of eight starting system configurations on mMethyl surfaces. A) and B) show the rotations by 0 and 180 degrees of $FnIII_{9\text{-}10}$ (rotations 90 and 270 are not shown). C) and D) show the rotations by 0 and 180 of $FnIII_{10}$ (rotations 90 and 270 are not shown). Corresponding initial system configurations were generated for mEA and mMA. Water molecules and ions are not shown. The PBC cell is shown as green vectors. The protein at the top is cut in half across the PBC box.

The new molecules were parameterised with CGenFF [140] which generated parameters for the functional groups. The new parameters for the forcefield CHARMM36m were the same as the parameters generated for the CHARMM36 in the previous chapters. For each of the functional groups, two SAM systems of different size, one 79 Å x 79 Å and the larger size was 107 Å x 107 Å, were assembled and solvated with GROMACS tools. The mEA SAM is illustrated in Figure 8.1B. For the equilibration of the surface, 50 Å of water was used between the surfaces in the z-dimension.

The system minimisation followed the protocol described in Chapter 3. The equilibration was carried out using a 100 ps simulation with the NVT ensemble at the temperatures of 200K and then 300K in both which the Nosé-Hoover thermostat was applied every 1 ps. After that, the NPT ensemble was used in a 100 ns simulation in which a Parrinello-Rahman semiisotropic barostat was employed with the reference pressure of 1 bar in the z-dimension. Throughout the process the sulphur atoms were restrained with 10,000 kJ mol^{-1} nm^{-2} in each dimension with respect to the original position. This restraint is necessary due to the absence of the gold

slab that previously held the sulphur atoms in place. All atoms were a subject to reference coordinates scaling due to pressure coupling.

The cut-off distance for the Coulomb and the van der Waals potential was set to 12 Å, and the Particle Mesh Ewald (PME) was used to calculate the long-range electrostatic interactions. The leap-frog integrator along with the Verlet cutoff scheme was used throughout all the simulations.

The equilibrated tilt and roughness of the chains are shown in the Table 8.1. To calculate the tilt, a vector from the sulphur to the carbon preceding the functional group was used. Then, an angle between this vector and the normal of the surface was calculated. In comparison to the SAMs on the gold slab, on average, the tilt has increased slightly for each new system, whereas the roughness increased for EA only. The new systems have the advantage that they present the same surface in the 3D PBC up and down. This way the protein can approach the surface from either side. Furthermore, the removal of the gold slab and the walls in the z-dimension further improves the performance significantly.

For the larger surfaces, two domains FnIII$_{9\text{-}10}$ were rotated four times to create four different orientations of the protein, called r0, r90, r180 and r270. The number in the replica name corresponds to the degree by which the protein fragment was rotated around its longest principal axis. For the smaller surfaces, four replicas (r0-r270) were created to study the adsorption of the FnIII$_{10}$ domain, in which the 10$^{\text{th}}$ domain was rotated the same way.

After that, in each of the systems, the protein was inserted to be the same distance away from the SAMs in the z dimension across the periodic image. After insertion of the protein, the water molecules which overlapped with any of the protein atoms within 2 Å were removed. The protein was modelled with CHARMM36m forcefield and the system was neutralised with a sodium ion when the FnIII$_9$ domain was present. Then, the previously described protocol consisting of energy minimisation, equilibration using the NVT ensemble at 200K and 300K, followed by a production simulation employing the NPT ensemble at 1 bar and 300K, was used. The sulphur atoms remained restrained. Each production simulation is 1 μs long, yielding 24μs of data.

For the calculation of the protein distance to the surface, only a subset of the heavy atoms

from the surface was used. From each chain the SAM sixteen atoms were used: eight on each end of the molecule.

8.1.1 Clustering

In order to define the adsorption states, clustering with the density-based spatial clustering of applications with noise (DBSCAN) algorithm was carried out. First, for each of the systems, the four replicas r0, r90, r180 and r270 were merged into a single trajectory. Then, for each residue, the smallest distance from its heavy atoms to the heavy atoms in the surface was calculated over time. Frames were used at the frequency of 1 ns resulting in 4 thousands frames for each system. The datasets were separated for the 9^{th} and 10^{th} domains. To compare any two frames, the root mean square deviation (RMSD) of the distances between the residues and the surface was calculated. The farthest distance that was allowed for any two frames to be in the same cluster was set to 1.7 Å. This distance generates coherent clusters while also finding all the adsorbed states in the datasets. This was investigates by visually checking the clusters and the frames labelled as noise. Then, DBSCAN was used on each dataset. During the analysis of each cluster, the position of the frames in the trajectories was investigated in order to establish the order from which cluster to which other cluster the simulation has moved on. Additionally, DBSCAN was set to have at least 50 different frames (equivalent to 50 nanoseconds) in order to create a cluster. Decreasing this number substantially did not increase the number of found clusters.

8.2 Adsorption

In this section the adsorption to the mMA, mEA and mMethyl SAMs with the updated force-field CHARMM36m is discussed. For each the surfaces, there are two sets of four replicas. The first set is used to investigate the adsorption of the tandem $\text{FnIII}_{9\text{-}10}$ whereas the second set contains only the FnIII_{10} domain. This is also the order in which the simulations are discussed in mEA and mMethyl SAMs.

8.2.1 mMA SAMs

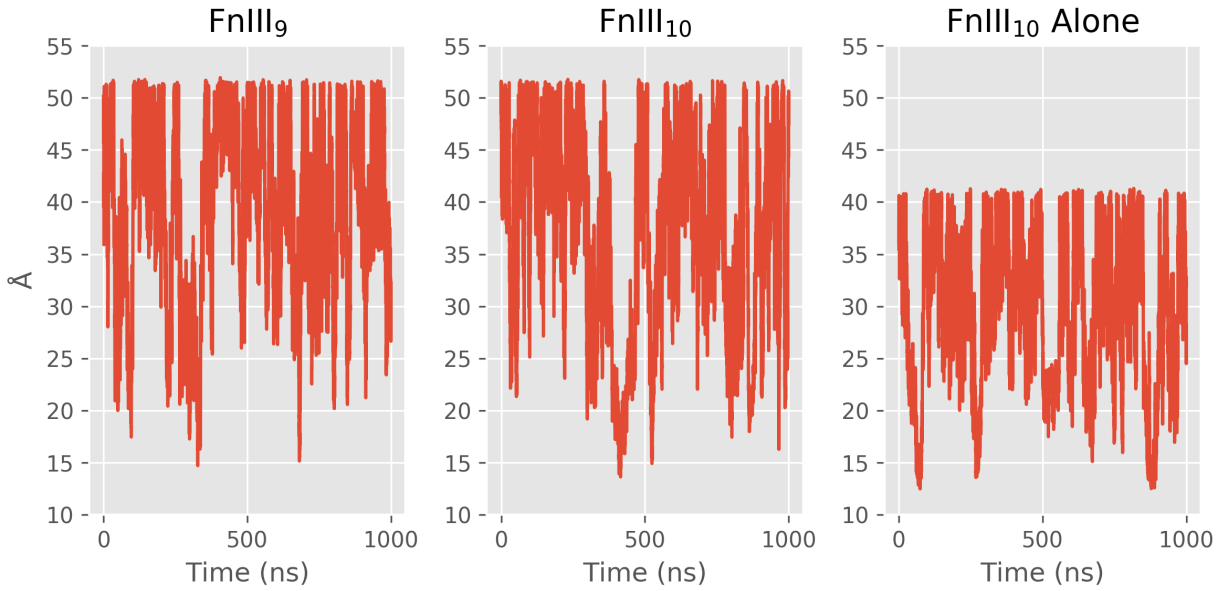


Figure 8.3: The distance from the centres-of-mass of each domain to the nearest heavy atom in the mMa SAMs in the r0 orientation. The other orientation r90-r270 follow the same pattern.

The distance from the centres-of-mass of each domain to the nearest heavy atom in the surface was calculated and the results for the mMA FnIII$_{9\text{-}10}$ and mMA FnIII$_{10}$ are shown in Figure 8.3. Only the orientation r0 is shown because the other orientations follow the same pattern. In the case of the tandem simulations, neither the 9$^{\text{th}}$ nor the 10$^{\text{th}}$ domain at no point remain close to the surface. This is despite the fact that each of them approaches the surface on multiple occasions in each of the replicas. The 10$^{\text{th}}$ domain simulated by itself similarly never stays close to the surface. Therefore, no adsorption is observed in any of the replicas to the mMethyl surface. For this reason no further adsorption analysis was carried out for the mMA surface.

8.2.2 mEA SAMs

FnIII$_{9\text{-}10}$ The nearest distance from the centres-of-mass of each domain to the heavy atoms in the surface was calculated for the mEA SAMs surface. The results are presented in Figure 8.4, with the results for the 9$^{\text{th}}$ and 10$^{\text{th}}$ domains plotted separately. In contrast to mMA SAMs, adsorption is observed to mEA SAMs, and therefore all four replicas are shown.

The 9$^{\text{th}}$ domain never fully adsorbs to the surface. The only contact it makes is in the first

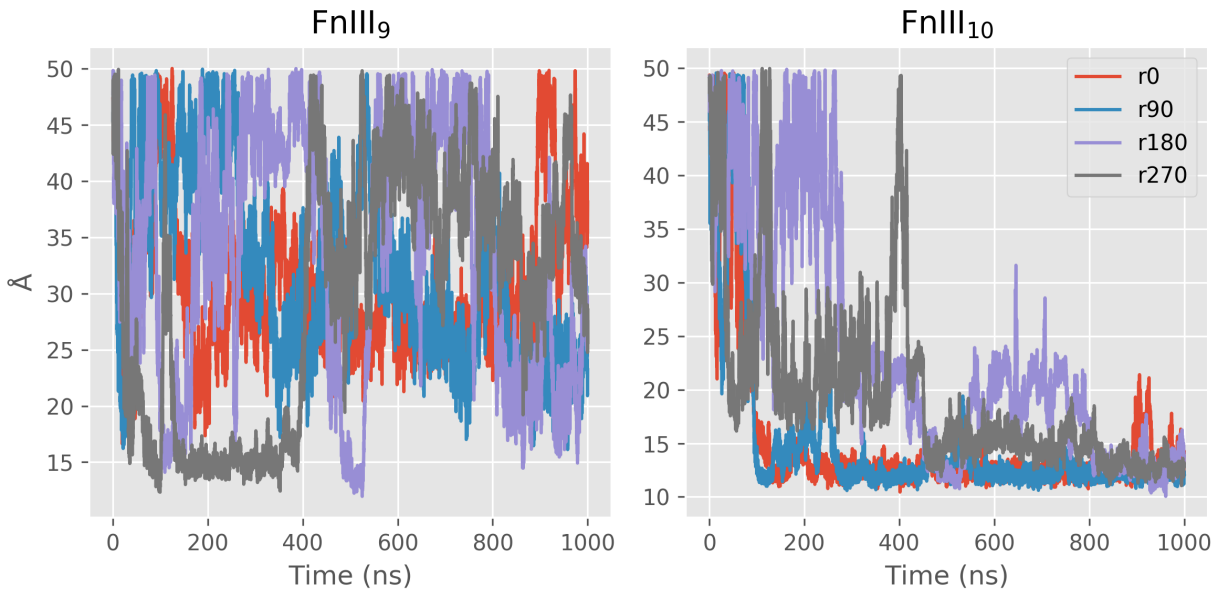


Figure 8.4: The distance from the centres-of-mass of each domain to the nearest heavy atom in the EA SAM. This is for all the four different rotations of the two domains.

quarter of the r270 replica, where it fluctuates steadily 15 Å away from the surface. However, even then it leaves without returning to the surface.

Out of the two domains, the 10^{th} domain adsorbs more strongly to the mEA SAMs. In the r0 and r90 replicas, the domain approaches the surface at around 150 ns time and then stays adsorbed. However, short fluctuations are seen in the r0 system towards the end, although the contact is not lost. In the other two replicas, r180 and r270, the 10^{th} domain does not adsorb to the surface until much later, shortly after 400 ns time. This suggests that these two protein orientations make it more difficult for the 10^{th} domain to adsorb. In these two replicas, the 10^{th} domain continues to approach the surface closer, finally converging with the systems r0 and r90 after 800 ns time.

Residues To understand which different adsorption states take place during the contact made by the 9^{th} domain and in the adsorption of the 10^{th} domain I used the distances from each residue to the nearest heavy atom in the surface SAM over time. Clustering with DBSCAN was carried out for the distances in the 9^{th} and 10^{th} domains separately, as described in Section 8.1.1. The computed clusters with different adsorption states for each domain are shown in Figure 8.5.

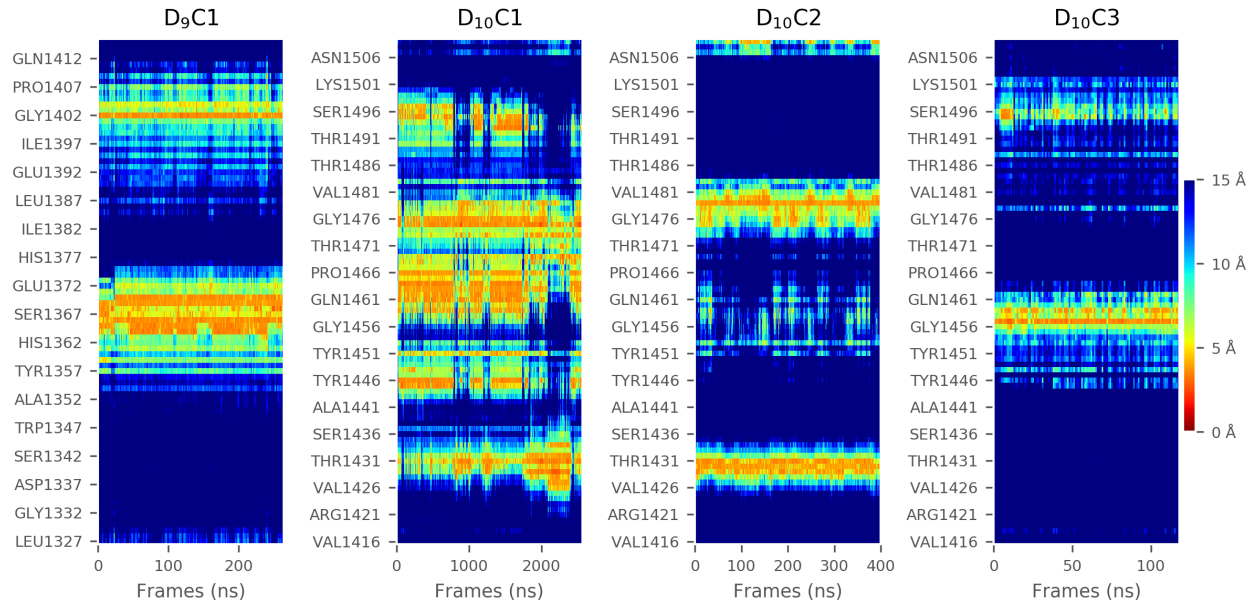


Figure 8.5: The cluster found in the adsorption across the four mEA$_{9\text{-}10}$ r0-r270 systems. For details of the analysis please see section 8.1.1. One cluster was found for the 9$^{\text{th}}$ domains and three clusters were found for the 10$^{\text{th}}$ domain. The X-axis are different and depend on the size of the cluster. Each frame represents 1 ns.

One adsorption state, labelled D$_9$C1, was found for the 9$^{\text{th}}$ domain and it contains 264 frames (equivalent to 264 ns). Out of all the clustered frames, 245 frames were found in r270, which is the only replica where the 9$^{\text{th}}$ domain makes a clear contact. The other frames were found in r180 (19 frames). The adsorption state D$_9$C1 has two adsorbing residue regions: one large and one small. The large region surrounds five residues which are around 4 Å away from the surface: Glu1364, His1365, Phe1366, Arg1369, Pro1370. These residues reside on the flexible loop$_{\text{C/D}}$ on the 9$^{\text{th}}$ domain. The smaller region contains a single residue that is consistently close to the surface, Gly1402, which is on a shorter loop on the same face.

During the adsorption of the 10$^{\text{th}}$ domain there are three adsorption states which are labelled D$_{10}$C1-3. The D$_{10}$C1 cluster is the largest one and contains 2566 frames together and is found in most of the replicas r0 (876 frames) and r90 (862 frames), as well as in r180 (308 frames) and r270 (517 frames). Despite the classification by DBSCAN, the cluster does not have perfectly uniform adsorption and can be divided into two halves. The first half finishes at the 1738$^{\text{th}}$ frame of the cluster, which represents the two replicas r0 and r90. After that frame, the second half comes from r180 and r270 and converges with the first half towards the end of the time

frame (2480 - 2560 frames). This transition of the two replicas r180 and r270 indicates that the first half is the more stable part of the cluster, whereas the second half is an intermediate state.

In both halves of $D_{10}C1$ two regions of residues always remain close to the surface. The first region surrounds Thr1431 and spans more residues in r180 and r270 with as many as eight residues below 5 Å distance away from the SAMs at one point. Among the eight residues, only one residue adsorbs as well as Thr1431, that is Thr1429. The second region surrounds Ser1475 with several residues close the surface, where Gly1476 is particularly close. This residue is not close to the surface during the intermediate part of the cluster. These residues, particularly close to the surface at all times, are threonine and serine, which are spatially next to each other, and which both have hydroxyl groups. Therefore, residues Thr1429, Thr1431 and Ser1475 potentially function as an anchor for the 10^{th} domain. In the simulations of $FnIII_{8\text{-}10}$ on methyl-terminated SAMs carried out by another group the residue Thr1431 was also shown to have the function of an anchoring residue [93].

Furthermore, in the $D_{10}C1$ adsorption state, several other residue regions can be distinguished. The first region surrounds Asp1495 which belongs to the RGD motif. However, the frequent diffusion seen in this residue region indicates that the motif does not directly interact with the surface, and I suggest that it is the flexibility of the loop F/G that leads to this state (see Figure 1.2). The largest residue region spans residues Gln1461, Glu1462, Phe1463, Thr1464 and Pro1466. These residues reside on the β - strand D just next to the interacting loop (Stand names are shown in Figure 1.2).

Two other residues can be distinguished for their distance frequently being less than 5 Å to the surface: Tyr1451 and Tyr1446. Interestingly, tyrosine also contains a hydroxyl group. Together with the previously discussed threonine and serine, there are now five residues altogether that are consistently close to the surface and which contain a hydroxyl group. It is plausible that the hydroxyl group aids adsorption via the water-mediated interactions with the surface.

The second adsorption state $D_{10}C2$ is found largely in r180 containing 320 frames. It has two residue regions adsorbed to the surface. The first region relies on residues Thr1429, Pro1430

and Thr1431 whereas the second region largely surrounds Pro1479. A further inspection of the adsorption state in the r180 replica shows that it transitions one time from $D_{10}C1$ to $D_{10}C2$, but eventually returns to the adsorption state $D_{10}C1$. Therefore, I classify the adsorption state $D_{10}C2$ as an intermediate state resulting from the different initial orientation of the two domains FnIII$_{9\text{-}10}$.

The last adsorption state, $D_{10}C3$ is found only in the r270 replica and counts 118 frames altogether. It involves one region consistently that focuses on a single residue Asn1457. This adsorption state is present while the 9th domain makes its one contact (adsorption state D_9C1), and quickly transitions to the adsorption state $D_{10}C1$ making it also an intermediate state.

The main adsorption state $D_{10}C1$ dominates the adsorption, regardless of the initial orientation of the two domains, which only add intermediate states. Although the 9th domain makes contact in the r270 replica, it loses it when the 10th domain adsorbs. Therefore, in that replica, the 10th domain dominates the adsorption too.

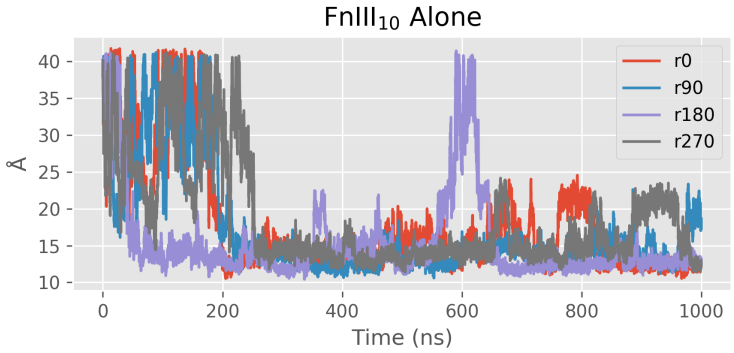


Figure 8.6: The distance from the centre-of-mass of the 10th domain, simulated by itself, to the nearest heavy atom in the SAM.

FnIII$_{10}$ Alone The distance from the centre-of-mass of the 10th domain, simulated by itself, to the surface was calculated for each replica. The results are presented in Figure 8.6. The domain adsorbs in all four replicas as indicated by the distance being mostly below 15 Å away from the surface. However, the distance to the surface does not always remain steady, and instead fluctuates in each of the replicas.

In addition to the observed fluctuations across the systems, in the mEA replica r180, the 10th

domain diffuses away from the surface for approximately 40 ns just before reaching t = 600 ns of the simulation. However, after readsorption to the surface at around 630 ns, the domain appears to be more stably adsorbed, with the distance fluctuating around 13 Å away from the surface. In the other replicas, r0, r90 and r270, the 10^{th} domain appears to regularly experience a weaker binding where the distance fluctuates around 20 Å away from the surface, only to return to the distance of 12 Å. One example can be found in the replica r0 with two periods where the domain moves to around 20 Å away from the surface starting at 650 ns and then at 750 ns. However, this behaviour is also observed in the r270 replica and towards the end of the r90 replica.

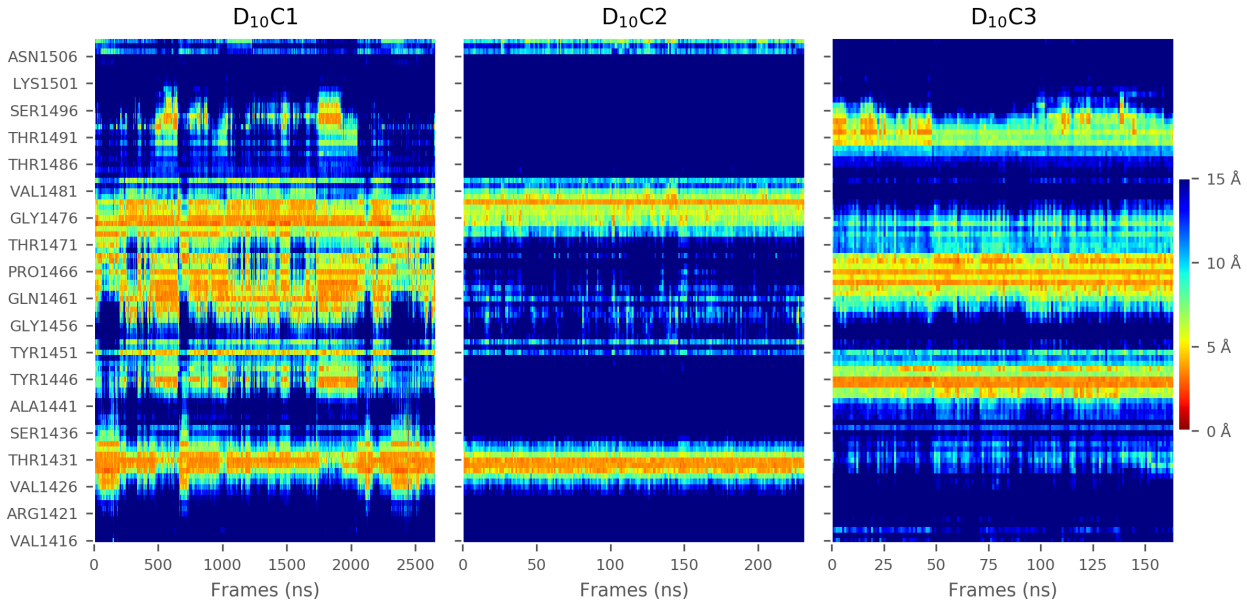


Figure 8.7: The clusters found in the adsorption across the four mEA_{10} r0-r270 systems. For details of the analysis see section 8.1.1. For the 10^{th} domain three adsorption states were found. The X-axis differ and depend on the size of the cluster.

Residues As described above, DBSCAN was used to find the adsorption states across the four replicas of the differently rotated $FnIII_{10}$ domain (See Figure 8.7). Three clusters were found with the major adsorption state $D_{10}C1$ containing 2663 frames spread relatively equally across the four replicas.

The largest adsorption state, $D_{10}C1$, shows some variations but some residue regions adsorb consistently. The first region contains residues Thr1429, Pro1430, Thr1431 and often Ser1432.

It is the same region found to be of importance during the adsorption of tandem $FnIII_{9\text{-}10}$ on the same SAMs. The second region spans residues 1461-1469 which reside on the β - strand D and contain Gln1461, Phe1463, Thr1464 and Pro1466. The third region surrounds Ser1475 and includes Thr1473, Ser1475 and Gly1476. Moreover, the residues Tyr1446 and Tyr1451 are found close to the surface without any other surrounding residues. These same residue regions were involved in the major adsorption state $D_{10}C1$ during the adsorption of tandem $FnIII_{9\text{-}10}$ to mEA, which means that the 10^{th} domain adsorbs the same way, regardless of whether it is alone or whether it is linked to the 9^{th} domain.

The second cluster contains 232 frames and is present mostly in the r0 and r270 replicas. This cluster corresponds to $D_{10}C2$ in the tandem $FnIII_{9\text{-}10}$. The last cluster has together 164 frames and is found mostly in r90. This adsorption state takes place midway through the replica which splits the main cluster $D_{10}C1$ in half, meaning that the $D_{10}C1$ ultimately replaces it. For this reason it is also an intermediate state. It is identified by two residue regions: first Tyr1446 and Arg1445 and second Thr1464 and Pro1466.

The 10^{th} domain, when simulated alone, adsorbs in a very similar fashion to when it is linked to the 9^{th} domain. It appears, however, less stably adsorbed, which is seen from the small fluctuations described with the centres-of-mass of the domain. Although the domain does not adsorb well in the tandem simulations in the r180 and r270 replicas, these are the two replicas which take longer to converge to the main $D_{10}C1$ adsorption state. In these two replicas the 10^{th} domain continues improving leading to stable adsorption after 800 ns. Therefore, it appears that the 10^{th} domain is more stably adsorbed in the presence of the 9^{th} domain. Despite the less stable adsorption of the 10^{th} domain alone, with the exception of the r180 replica, the domain does not leave the surface. Moreover, the often present hydroxyl-group-containing residues such as tyrosine and threonine also reappear, which suggests that they play an active role in the adsorption.

8.2.3 mMethyl SAMs

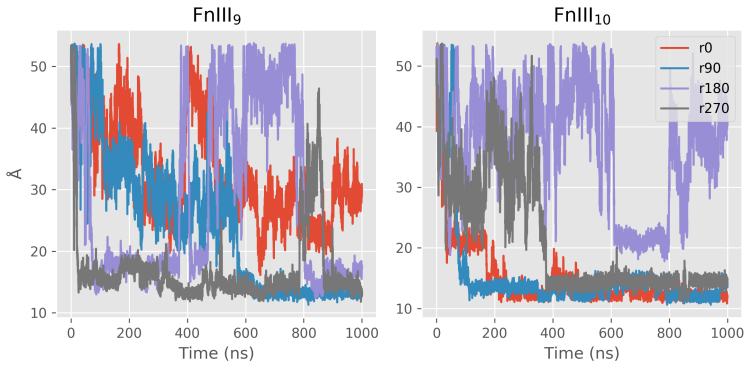


Figure 8.8: The distance from the centres-of-mass of each domain to the nearest heavy atom in the SAM.

FnIII$_{9\text{-}10}$ The distance from the centres-of-mass of both domains to the nearest heavy atom in the surface was calculated. The results are shown in Figure 8.8 and represent the overall adsorption trends seen on the mMethyl SAMs. In the replica r0, the 9th does not adsorb, but the 10th domain does. The 10th domain approaches the surface at 35 ns and its adsorption rapidly improves at 170 ns time. Except for small fluctuations, it largely remains at 12 Å away from the surface until the end of the simulation.

In the r90 replica, the 9th domain makes first contact at 580 ns time and then comes closer to the surface, fluctuating around 13 Å away from the surface for the rest of the simulation. In this replica, the 10th domain also adsorbs well, starting from 80 ns time and then either fluctuating either 13.5 Å or 12.5 Å away from the surface for the rest of the simulation.

The r180 replica follows a different pattern and adsorbs only weakly. The 9th domain makes contact at 180 ns time where it stays poorly adsorbed (average distance 18 Å) until 370 ns, which is followed by diffusion from the surface for 40 ns. The domain makes another contact between 800 ns and 1000 ns time, but it never steadily adsorbs. Similarly, in this replica the 10th domain only makes good contact between 610 ns and 800 ns.

In the r270 replica the pattern changes once again. This time the 9th domain makes contact early at 25 ns time and then remains around 13 - 16 Å away from the surface. The major exception to this is the period between 780 ns and 880 ns where the domain is detached from the surface. The 10th domain contacts the surface at 380 ns and then remains 14 Å away from the surface.

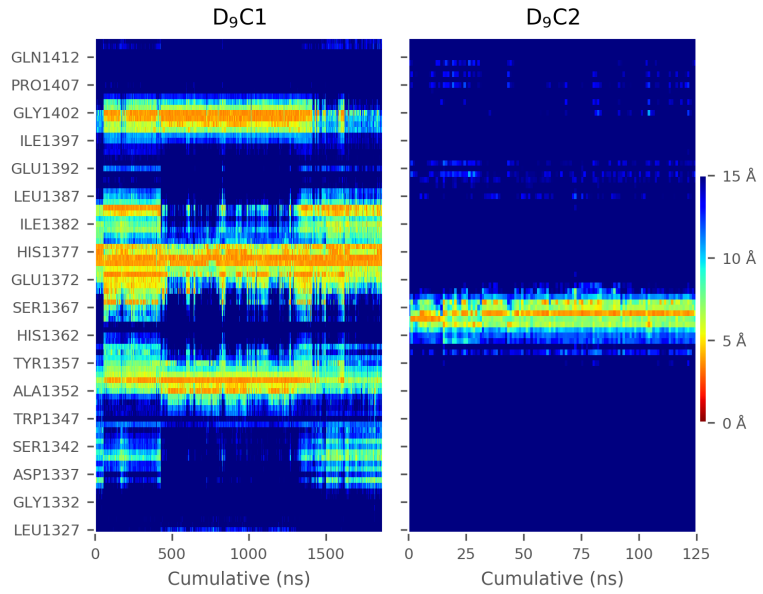


Figure 8.9: The two adsorption states of the 9$^{\text{th}}$ domain across the r0 - r270 replicas. For details of the analysis please see section 8.1.1. The X-axis differs depending on the size of each cluster.

Residues DBSCAN clustering was used to find similar adsorption states across the four replicas with differently rotated FnIII$_{9\text{-}10}$ protein fragment, as described in section 8.1.1. Altogether, there are two adsorption states assumed by the 9$^{\text{th}}$ domain and three adsorption states assumed by the 10$^{\text{th}}$ domain.

The first adsorption state of the 9$^{\text{th}}$ domain (Figure 8.9, D$_9$C1) has 1872 ns and is found spread uniformly across the replicas r90 - r270. It contains one residue region that is present at all times which is centred around Val1375 and Pro1376. The other residues in the region include Asp1373 and Ser1378. The second region that is close to the surface for most of the simulation time is centred around Ile1354. This site also involves Ala1352 making it a hydrophobic spot. The third region has two residues particularly close to the surface: Asn1401 and Gly1402. This site also has two hydrophobic residues Ala1399 and Leu1400, which, however, are on average slightly farther away from the surface than the Asn1401 and Gly1402 residues. In addition to the aforementioned regions there are two single residues which are close to the surface at different points: Leu1384 and Thr1385.

The second adsorption state D$_9$C2 contains 125 frames altogether and is found only in the r0 replica. It uses one residue area that relies mostly on the residue Phe1366. This contact takes

place when the 10th domain is well adsorbed.

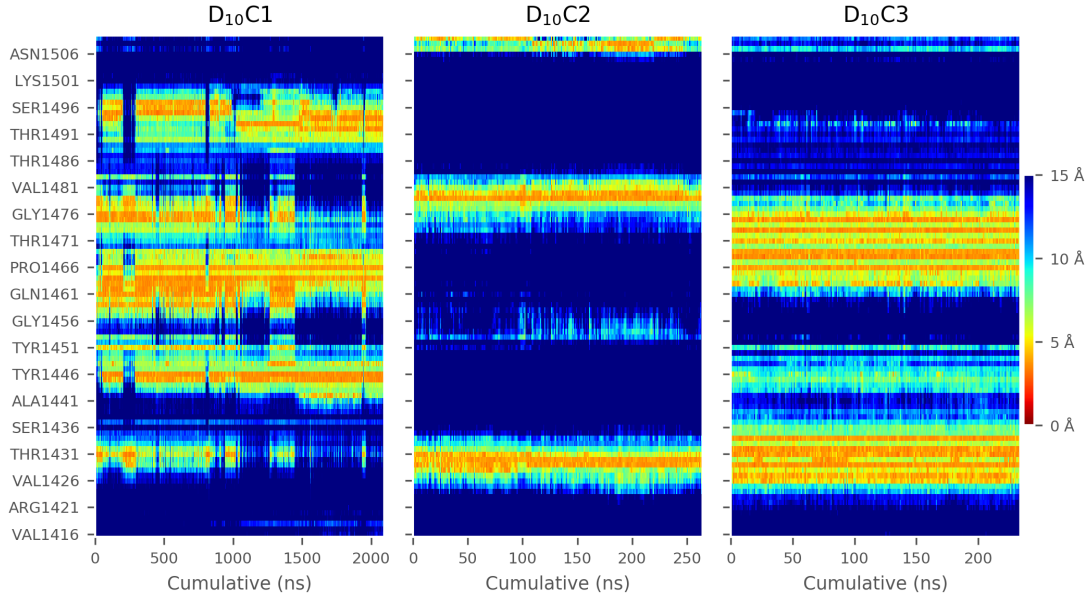


Figure 8.10: The three clusters in the adsorption of the 10th domain across the mMethyl$_{9\text{-}10}$ r0-r270 replicas. For details of the analysis please see the methods section 8.1.1. The X-axis differ and depend on the size of the cluster.

There are three adsorption states of the 10th domain, with D$_{10}$C1 counting 2096 ns altogether. The cluster has three residue regions which remain in contact with the surface most of the time. The largest residue region is centred around Thr1464 and Pro1466, with Val1465 in between them. Some of the residues in the region, 1459-1461, lose their binding, whereas Glu1462 and Phe1463 remain close to the surface which is likely due to the hydrophobicity of Phe1463. The second region that remains close to the surface has only two residues which are Arg1445 and Tyr1446. The third region is close to the N-terminal and appears to be able to adsorb in two different ways. During the first half it uses Asp1495, Ser1496 and Pro1497, and in the second half the residues Gly1492, Arg1493 and Gly1494 are close to the surface. In addition to these regions, there are also the two residues Ser1475 and Gly1476 which are consistently close to the surface during the first half of the adsorption state.

The other two adsorption states D$_{10}$C2 and D$_{10}$C3 are smaller containing only 264 and 234 frames (ns), respectively. The D$_{10}$C2 is split between r0 and r180 and involves significantly fewer residues than the main cluster D$_{10}$C1. It has two main regions: one with Thr1429 and Pro1430, and the other surrounds Pro1479. The position of the frames in this adsorption state

shows that $D_{10}C2$ is an intermediate state in r0 and r180. In the latter, it is followed by the loss of adsorption of the 10th domain and the adsorption of the 9th domain.

The last cluster $D_{10}C3$ is found only in the r90 replica. It is found near the beginning of the replica and is lost to the $D_{10}C1$ cluster, making it a intermediate state. It has two large residue regions involved in the adsorption. The first region spans residues Val1426 to Leu1434 where Thr1429, Thr1431, Ser1432 are worth mentioning for being consistently close to the surface. The second region spans just as many residues starting from Pro1466 to Ser1475, where the residues often close to the surface include Ser1468, Lys1469, Thr1471 and Thr1473.

Across the four replicas the tandem FnIII$_{9\text{-}10}$ adsorbs in different ways, with both domains having their own dominant adsorption states each counting more than 1500 ns. However, the two adsorption states D_9C1 and $D_{10}C1$ are not exclusive and coexist in the replica r90 and r270.

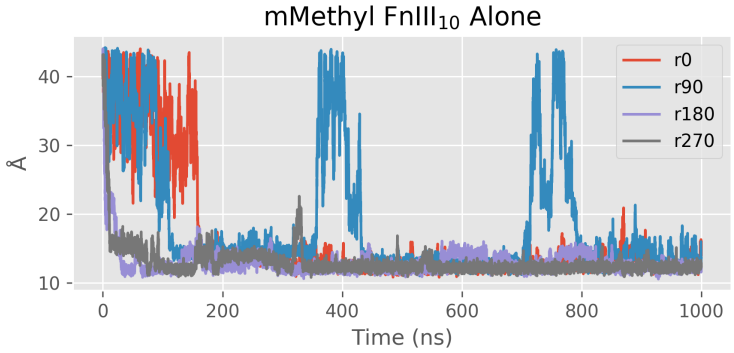


Figure 8.11: The distance from the centre-of-mass of the 10$_{th}$ domain to the nearest heavy atom in the methyl SAMs.

FnIII$_{10}$ Alone The adsorption of the 10th domain is described by the distance from the domain's centre-of-mass to the nearest heavy atom in the methyl SAMs (Figure 8.11). In each of the four replicas the 10th domain has adsorbed to the surface. However, in the r90 replica, the domain diffuses away from the surface twice: at 360 ns time for 60 ns and at 720 ns time for 70 ns. A similar event is noted in the r270 replica where at 315 ns time the distance increases to 22 Å from the surface. Although during this period the N-terminal is still in touch with the surface.

Whereas in the other replicas the domain does not move away from the surface to the same extent, it appears to transition to the distance of 14 - 15 Å and then back to 12 Å away from the surface. This happens on multiple occasions in the r0 and r180 replicas. The general trend is, however, that once the 10th domain makes contact, it stays close the surface.

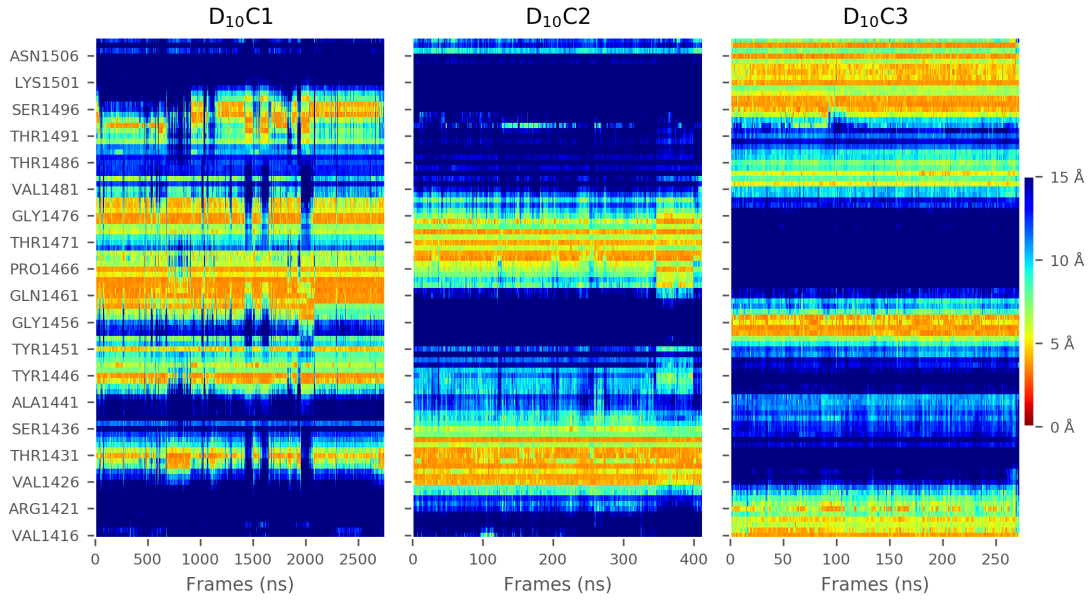


Figure 8.12: The adsorption states of the 10th domain to the mMethyl$_{10}$ SAMs across the r0-r270 replicas. For details of the clustering please see the section 8.1.1. The X-axis differ and depend on the size of a cluster.

Residues DBSCAN was used to find the common adsorption patterns across the four replicas. Three different clusters were found (Figure 8.12) with the adsorption state, D$_{10}$C1, counting 2757 frames altogether. Although the frames are present in every replica, they are more rare in the r90 replica (203 frames).

The adsorption state D$_{10}$C1 involves five different residue regions. The largest is centred on Phe1463 surrounded by Gln1461, Gln1462, Thr1464 and Pro1466, which are consistently close to the surface. This region was also recognised in the adsorption of the 10th domain and in the adsorption of the tandem domains to mMethyl SAMs. Furthermore, this adsorption state corresponds closely to the adsorption state D$_{10}$C1 from the corresponding tandem simulations. There is a small difference in that the adsorption state of the 10th domain by itself is more consistent - the adsorbed regions do not diffuse from the surface as frequently. This suggests

that the 9th domain interferes in the adsorption of the 10th domain, which is consistent with the finding that the 9th domain has its own independent adsorption state, as discussed in the previous section.

The second cluster D$_{10}$C2 contains frames mostly from the r0 and r90 replicas. This is an intermediate state that takes place during the initial contacts. Furthermore, this adsorption state has previously been found in the adsorption of the tandem to the mMethyl SAMs and was labelled D$_{10}$C3.

The third adsorption state is not present in the adsorption of the tandem to mMethyl SAMs. All of its 273 frames come from the replica r90. This adsorption state is seen in the middle of the replica and is followed by the 10th domain diffusing from the surface. After that, it returns reoriented and adsorbs using the adsorption state D$_{10}$C1; the state in which it remains until the end of the replica.

Thus, the 9th domain stably adsorbs to the hydrophobic surface as represented by the D$_9$C1 adsorption state, whereas the 10th domain relies largely on the same residues during its adsorption, regardless of whether it is attached to the 9th domain or not.

8.3 RGD and PHSRN Motifs

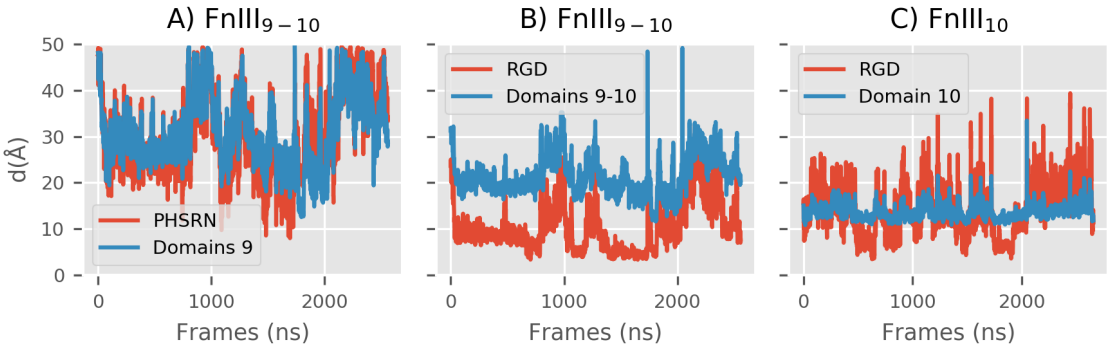


Figure 8.13: The binding availability of the motifs. For **A)** and **B)** the analysis is carried out on the main adsorption state D$_{10}$C1. For **C)** the RGD motif is presented for the D$_{10}$C1 on mEA when the FnIII$_{10}$ domain is alone.

In the previous sections I discussed the major adsorption states of the 9th and 10th domains

on mEA and mMethyl SAMs. Here, for each of the main adsorption states, the availability of the two motifs, RGD and PHSRN is measured. Specifically, the distance from the centre-of-mass of each motif is measured to the nearest heavy atom in the surface. The PHSRN motif's distance is compared to a centre-of-mass of the 9th domain, and the RGD motif's distance is compared to a centre-of-mass of the two domains in the tandem FnIII$_{9\text{-}10}$ simulations, and to a centre-of-mass of the 10th domain when FnIII$_{10}$ is simulated alone.

Let us go back to the mEA surface and analyse the exposure of the two motifs in the main clusters D$_{10}$C1 in the tandem FnIII$_{9\text{-}10}$ and the single domain FnIII$_{10}$ simulations. The results are presented in Figure 8.13. The PHSRN motif follows closely the 9th domain in terms of the centre-of-mass distance to the surface throughout the adsorbed state. This means that the motif resides on the side of the adsorbed domain. The RGD motif is almost always closer to the surface than the centre-of-mass of the FnIII$_{9\text{-}10}$ domains. In other words, the RGD motif is not particularly available for binding. Furthermore, in the simulations with the 10th domain alone, the RGD-surface distance fluctuates significantly, sometimes being closer and sometimes farther from the surface than the centre-of-mass of the 10th domain. This is because the absence of the 9th domain makes it possible for the motif to move around more freely. This fluctuation means that the RGD residue is on the side of the adsorbed domain, which makes it more available for binding than the availability of the RGD motif in the tandem FnIII$_{9\text{-}10}$.

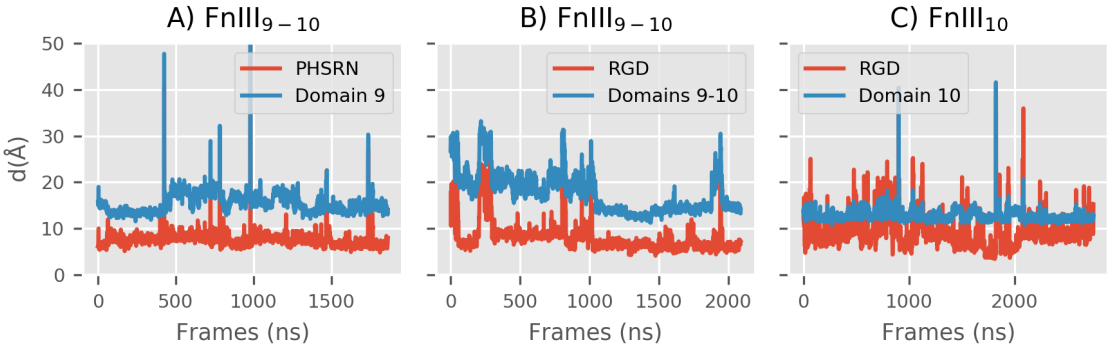


Figure 8.14: The binding availability of the motifs. In **A)** the distances were measured during the adsorption state D$_9$C1 when the 9th domain dominates the adsorption. In **B)** the motifs are presented for the D$_{10}$C1 when the FnIII$_{10}$ domain dominates the adsorption. In **C)** the RGD is presented when the 10th domain is simulated alone (D$_{10}$C1 cluster).

The availability of the motifs during the adsorption to the mMethyl SAMs follows a similar

pattern. When the 9th domain dominates adsorption, PHSRN appears to be consistently closer to the surface than the 9th domain (Figure 8.14A). This means that the motif might be partly trapped between the domain and the surface, or interact with the surface, or both. When the 10th domain dominates the adsorption, the RGD motif is consistently closer to the surface as well. Furthermore, during the simulation of the FnIII$_{10}$ alone, the RGD motif is also often closer to the surface than the the centre-of-mass of the 10th domain. Although there are fluctuations in the distance, the motif remains on average closer to the surface.

In general, both motifs are not very available for binding, particularly during the tandem FnIII$_{9\text{-}10}$ adsorption. The effects are consistent across the two surfaces mEA and mMethyl SAMs. The RGD motif might be slightly more available for binding when the 10th domain is simulated by itself, however, even then, it is mostly found close to the surface.

8.4 Interdomain Orientation

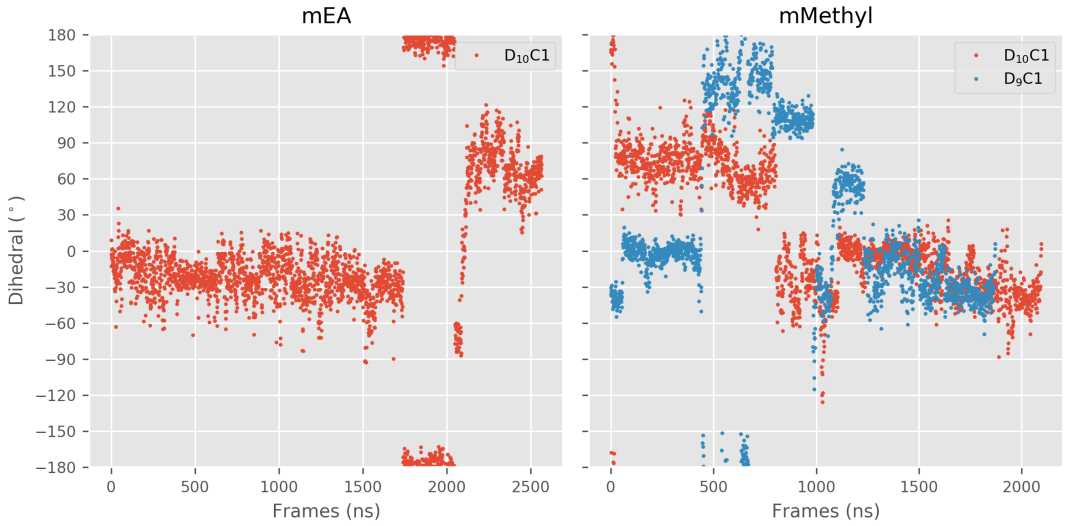


Figure 8.15: The interdomain orientation of FnIII$_{9\text{-}10}$ quantified with a super-dihedral (defined in Chapter 7). For mEA surface, the dihedral angle is presented during the adsorption state D$_{10}$C1. For mMethyl, the dihedral is presented when the 9th domain dominates the adsorption (D$_9$C1 cluster) and when the 10th domain dominates the adsorption (D$_{10}$C1 cluster).

In order to quantify the interdomain orientation of the two domains FnIII$_{9\text{-}10}$, the dihedral angle between the centres-of-mass of four residues is measured. The results are presented in Figure 8.15 for the largest adsorption states that I defined previously.

During the main adsorption state $D_{10}C1$ to the mEA surface, the dihedral angle mostly assumes three different angles. The most common angle fluctuates around -30°, which occurs in approximately 1700 frames. The second most common angle is -180/180° which occurs in approximately 350 frames. The third population of angles fluctuates around 70° for around 400 frames. It appears that when the 10^{th} domain dominates the adsorption, it does not force a particular interdomain orientation on the 9^{th} domain.

The interdomain orientation on the mMethyl substrate is different depending on whether the 9^{th} (D_9C1 cluster) or the 10^{th} domain ($D_{10}C1$) dominates the adsorption. When the 9^{th} domain drives the adsorption, several different dihedral angles have been observed, with the range 10° to -40° being the most common one. The next two dihedral angles are seen in the replica r180 and fluctuate around 145° and then 110° for almost 500 ns altogether. The dihedral angle 55° spans almost 200 frames, but in the same replica, r270, it changes to the most common range 10° to -40°.

Similar dihedral angles have been observed when the 10^{th} domain drives the adsorption. Two main dihedral angles can be distinguished: 70° for around 750 frames and -30° for over 1000 frames. Therefore, it appears that regardless of which domain drives the adsorption, the most common dihedral angle is found spanning the range 0° to 40°, which contains the dihedral angle of the original crystal structure (0°, PDB:1FNF).

The two adsorption states on the mMethyl SAMs are not exclusive but they describe most of the simulation time. The two main adsorption states of the two domains often take place at the same time. This overlap is present in two replicas: r90 and r270. A further look at the dihedral angle in the two replicas shows that in the r90 replica, during the overlap, the dihedral is fluctuating around 0°. In r270, the situation is the same: it fluctuates around -45°. In other words, when the two domains are in their main adsorptions states at the same time, they also have the same dihedral angle that is found in the crystal structure (PDB:1FNF), which is 0°, and which also belongs to the same minima found in Chapter 7.

8.5 Discussion

In this chapter I simplified the three surfaces EA, MA and methyl SAMs by removing the gold slab and calling the new surfaces mEA, mMA and mMethyl SAMs. Furthermore, in the context of the unexpected partial unfolding of the 9$^{\text{th}}$ domain discussed in Chapter 6, I updated to the forcefield CHARMM36m, a successor of CHARMM36. The new forcefield is in better agreement with the experimental structure in terms of the interface between the 9$^{\text{th}}$ and the 10$^{\text{th}}$ domain (Chapter 7). With the newly modelled surface and newer forcefield, I ran a set simulations for the FnIII$_{9\text{-}10}$ and FnIII$_{10}$ protein fragments with four different starting orientations.

mMA & MA Neither the tandem FnIII$_{9\text{-}10}$ or FnIII$_{10}$ make contacts with mMA surface across the eight simulations. In Chapter 3 I discussed simulations of the FnIII$_{9\text{-}10}$ domains on the MA SAMs. There was no adsorption but contacts have been observed, followed by the protein diffing from the surface (Figure 3.2). Therefore, due to the lack of any contacts, the binding on mMA is even weaker than on MA SAMs, while the results are consistent.

mEA & EA On the mEA, which contains the tandem domains FnIII$_{9\text{-}10}$, it is the 10$^{\text{th}}$ domain that dominates the adsorption. The clustering algorithm DBSCAN found the main adsorption state, D$_{10}$C1, which describes the way the protein adsorbed. This adsorption state is reached regardless of the initial orientation of the protein. Furthermore, the 10$^{\text{th}}$ domain simulated by itself reaches the same adsorption state, which altogether shows a clear preference for the domain to adsorb in one way. The other found adsorption states are intermediate and are followed by D$_{10}$C1. In contrast to the 10$^{\text{th}}$ domain, the 9$^{\text{th}}$ domain makes only a single contact on mEA.

Previously, in Chapter 3, in the simulations of FnIII$_{9\text{-}10}$ on EA SAMs, it was the 9$^{\text{th}}$ domain that dominated the adsorption. Also, it was the 10$^{\text{th}}$ domain that continually kept trying to adsorb. This was complemented with simulations of the 10$^{\text{th}}$ domain by itself in Chapter 4 which showed that the domain adsorbs well to the surface. Here, with CHARMM36m and the simplified mEA surface it is the 10$^{\text{th}}$ domain that drives the adsorption. In fact, the 9$^{\text{th}}$ domain

adsorbs poorly to the surface.

Let us compare the adsorption of the 10^{th} domain to mEA and to EA which was described in Chapter 4. In EA10 the domain adsorbs the same way as the adsorption state $D_{10}C1$. One difference is that the residues in EA10 are on average closer to the surface. On EA18, however, the adsorption of the domain is similar to $D_{10}C2$ on mMethyl surface (10^{th} domain simulated alone). However, even then the simulation is progressing to $D_{10}C1$, which is observed in the middle of the EA18 replica, and then again towards the end.

The residues consistently close the surface are in Table 4.1 and they are very similar to the residues driving the adsorption to mEA. For example, Thr1429 and Thr1431 are present in both chapters. Similarly, the residue region 1461-1469 which spans the β - strand D is present in EA10 and to a lesser extend in EA18. And the third region singled out with Ser1475 with Thr1473, Gly1476, is also found in EA10 and EA18. In other words, the 10^{th} domain adsorbs largely in the same way across the two forcefields, when the 9^{th} domain is absent.

mMethyl & methyl The tandem $FnIII_{9\text{-}10}$ adsorbs well to mMethyl SAMs with two main adsorption states: one in which the 9^{th} domain dominates the adsorption, and another in which the 10^{th} does. However, these two states are not exclusive. In the simulations of the 10^{th} domain by itself, the 10^{th} domain adsorbs the same way, with the minor difference of fewer residue-surface distance fluctuations.

The adsorption of tandem $FnIII_{9\text{-}10}$ to methyl SAM was also seen in Chapter 5. The main difference is that the adsorption of the protein fragment is different between the methyl10 and methyl18 systems. Here, across the four different orientations (replicas r0 - r270), each of the domains always converges to its major adsorption state. There are several intermediate states but they are replaced by these main adsorption states. This difference could be due to the increased timescale of the simulations ($1\mu s$ vs 500 ns). Although there was no indication of any changes in the adsorption towards the end of the simulations in Chapter 5.

On methyl18 SAMs, the 9^{th} domain uses mostly two residue regions: Asn1401 - Arg1403, and Arg1374 - His1377. The former is found close to the mMethyl in this chapter - with Asn1401 and

Gly1402 close to the surface. The latter also overlaps with the main adsorption state D_9C1, where Asp1373, Val1375, Pro1376 and Ser1378 were found consistently close to the surface. Moreover, the lone Thr1355 on methyl SAMs is just next to Ile1354 and Ala1352 which are close to mMethyl SAMs. Therefore, the 9th domain in the methyl18 replica adsorbs in a very similar way to the main adsorption state D_9C1 here.

The 10th domain adsorbs in two different ways across the methyl10 and methyl8 replicas, with few common residues. These are Arg1445 and Tyr1446, and Thr1464. The first two residues are close to the surface in the $D_{10}C1$ adsorption state, with no other residues around. This makes them consistent adsorbers across the two forcefields on the hydrophobic surfaces. The latter site includes only Thr1464 and is found in the adsorption state $D_{10}C1$, which also comprises the centre of the largest residue region that is close to the surface.

Furthermore, Ser1475 is close to the surface in methyl10, and Asp1495-Pro1497 in methyl18. These are sites found in the $D_{10}C1$ adsorption state on mMethyl. We see therefore close differences in which residues tend to be close to the surface.

Adsorption to mEA & mMethyl The 9th domain made one contact on mEA but adsorbed stably on mMethyl. Despite the huge difference, the Gly1402 residue was found to be common present in both of these cases.

The 10th domain has a distinctive adsorption state on mEA and on mMethyl. Interestingly, the adsorption of the 10th domain to mEA and mMethyl is largely the same. Each of the four residue regions adsorbing to the surface are present across the two surfaces, with the largest residue region surrounding Pro1466 and Thr1464. Each region has a clear overlap of the same residues. Even the lone residues Tyr1451 is found often close to the surface across the two surfaces. Therefore, it appears that the mEA surface has similar properties to the mMethyl surface, and the differences are rather subtle. They are sufficient for the 9th domain to starts adsorbing to mMethyl surface, but similar enough for the 10th domain to adsorb the same way.

The initial orientations of the tandem FnIII9-10 called r0, r90, r180 and r270 due to the degree by which the fragments were rotated showed little effect on the final outcome. On the mEA, the

r0 and r90 clearly had their 10^{th} domain adsorb earlier and in a more similar manner. However, r180 and r270 also converged to these results, after, but significantly later in the simulations, after 750 ns showing the most similar states. On the second surface on which adsorption was seen, mMethyl, there is more behaviour and r180 is least similar to the other replicas, and shows the weakest adsorption in terms of the number of residues close to the surface.

As is the case with most classical atomistic simulations, these results suffer from a sampling problem, and should therefore be taken with a pinch of salt. Further work is needed on the quantification of uncertainty in the results, and the related concept of convergence in the adsorption, for which enhanced sampling methods should be investigated.

Motifs In terms of motif exposure, the adsorption of the $FnIII_{9\text{-}10}$ domains is very different. On EA, both RGD and PHSRN are available most of the time during the adsorption (Figure 3.6). However, the two motifs are not available for binding during the adsorption with the CHARMM36m forcefield. It appears that they are partly hidden between the surface and the centre-of-mass of the two domains (Figures 8.13, 8.14). When the 10^{th} domain is simulated alone, the RGD availability is similar to that of EA18 (Figures 4.5, 4.5), where the motif is close the surface during adsorption, and is therefore partly inaccessible. This is a major difference that is also of importance when using either the tandem $FnIII_{9\text{-}10}$ or $FnIII_{10}$ to make the motif available for cells.

The two motifs are not available for binding after the domain adsorbed to the mMethyl SAMs. For example, PHSRN is not available for binding during the main adsorption state of the 9^{th} domain (D9C1). Similarly, the RGD is not available for binding in the main adsorption state of the 10^{th} domain (D10C1). However, it is not always the case in every replica here that when one domain adsorbs well, the second does too. In the previous chapter the results were similar. The two motifs are rarely available during the adsorption on methyl SAMs with CHARMM36 (Figure 5.7), although the PHSRN motif resides on the side of the 9^{th} domain in the Methyl10 system. Overall, however, the motif exposure is rather low.

Interdomain Orientation The interdomain interface is more relaxed in the new forcefield CHARMM36m where the 10$^{\text{th}}$ domain is the dominant domain adsorbing to the surface - and often, the only one. Different dihedral angles were sampled during the adsorption to mEA and mMethyl, but the majority is still found to oscillate around 20°. This was found to be the major minima for the two domains to assume in Chapter 7. The other dihedral angles are often present despite the fact the dihedral angle was measured during the same adsorption state of each domain. In other words, the adsorption of one domain does not the other domain to assume any particular interdomain orientation.

The CHARMM36 D9/D10 stronger interface prohibits D10 from fully adsorbing in Chapter 3, which is due to the competition of D9-D10 with each other. Here, the CHARMM36m, with the water model having a deeper well and therefore stronger interactions, the two domains were more able to stay out of each other's ways, which made them behave more independently.

Starting Configuration Sampling In order to improve sampling, the initial orientations of FnIII$_{9\text{-}10}$ and FnIII$_{10}$ I used were different. However, other methods can be used in order to optimise the selection of the initial protein-surface configurations. One established approach is that of Brownian dynamics, where the protein fragment is rigid and the water molecules are not explicitly represented. This approach can be used to quickly find simplified adsorption states and their energies. These different initial configurations can in turn be used with the more refined models and atomistic simulations or other enhanced sampling methodologies [161].

Chapter 9

Conclusions

My PhD began with trying to understand how a very small change in a polymer surface makes fibronectin form biologically active networks or aggregates. This was the topic of Chapter 3 in which I modelled the polymers poly(ethyl acrylate) and poly(methyl acrylate) as EA and MA self-assembled monolayers (SAMs). Using molecular dynamics simulations I showed that the domains $FnIII_{9\text{-}10}$, with the forcefield CHARMM36 (C36), adsorbed to the EA SAMs but not to MA SAMs. The experimental work by the group of Manuel Salmerón-Sánchez showed that the EA and MA SAMs reproduced fibronectin behaviour observed on the original polymers poly(ethyl acrylate) and poly(methyl acrylate). Specifically, they showed that fibronectin formed networks on EA SAMs but not on MA SAMs. Therefore, the simulations suggest that adsorption could be a factor in fibronectin fibrillogenesis. Then I showed that the difference in adsorption is due to the different hydration of the EA and MA SAMs. Specifically, that the small difference of one less methylene bridge significantly increased the hydration of the MA SAMs. The functional group showed less freedom of movement and therefore formed a denser hydration layer. This is in contrast to EA SAMs where, due to their one extra methylene bridge, the terminal group was free to rotate, disturbing the hydration. This difference in hydration affected adsorption of the domains $FnIII_{9\text{-}10}$ in the simulations, where the forcefield CHARMM36 (C36) was used. Another interpretation is that the outer methyl group motility affects how the surface interacts with the protein. The fibronectin fragment adsorbed to EA SAMs but not to the more densely hydrated MA SAMs.

The adsorption simulations were reproduced in chapter 8 with the newly released forcefield CHARMM36m (C36m). These new simulations used a simplified surface which allowed for improved sampling by employing more and longer simulations. Therefore different initial orientations of the $FnIII_{9\text{-}10}$ and $FnIII_{10}$ were probed. The results of these simulations were fundamentally the same as the simulations with C36. Adsorption of both $FnIII_{9\text{-}10}$ and $FnIII_{10}$ was taking place on EA SAMs but not on MA SAMs. However, there were important differences between the forcefields. With C36 the adsorption to EA SAMs was driven by the 9^{th} domain while the 10^{th} domain did not adsorb. The opposite behaviour was observed with the C36m forcefield. Instead, the 10^{th} domain drove the adsorption, while the 9^{th} domain stayed away from the surface. Another difference is that the motifs RGD and PHSRN were available for integrin-binding after adsorption in C36, but with C36m they were largely buried in the surface.

The 10^{th} domain in $FnIII_{9\text{-}10}$ did not adsorb to EA SAMs with the C36 forcefield. However, with the same forcefield, in Chapter 4 the 10^{th} domain was simulated by itself and it adsorbed well. I compared this adsorption to the corresponding simulations with the C36m forcefield in Chapter 8. In both forcefields the domain adsorbed in largely the same way. With C36m, the 10^{th} domain adsorbed to the EA SAMs the same way, regardless of whether the 9^{th} domain was present or not. Despite this consistency in the adsorption of the 10^{th} domain across the two forcefields, the adsorption of the tandem $FnIII_{9\text{-}10}$ was driven by the 9^{th} domain in C36 and by the 10^{th} domain in C36m. In other words, the later forcefield C36m appears to have affected the two domains to different extents. For these reason, it would be of interest to investigate the adsorption of the 9^{th} domain to EA SAMs by itself to further understand how each of the domains is affected by the change in forcefields.

Another surface on which adsorption was studied with the two forcefields is the hydrophobic methyl SAM. With C36 both domains adsorbed quickly and remained steadily adsorbed. I found only a few small similarities across the two replicas. Specifically, despite the number of residues involved in the adsorption, there was only one site in common across the two replicas. Therefore it was concluded that the adsorption to the methyl SAMs is non-specific.

This non-specific adsorption was not reproduced with the C36m forcefield. Each of the two domains, whether in tandem $FnIII_{9\text{-}10}$ or just the $FnIII_{10}$, converged to the same adsorption state. It was found that the 9th domain shows similarities in the overall adsorption to one of the replicas with C36. In addition, the 10th domain also shares a couple of key adsorption residue sites to C36. However, the lack of convergence in adsorption to methyl SAMs with C36 means that two very different sides of $FnIII_{9\text{-}10}$ were involved in adsorption. Therefore, at least some of the similarities could be spurious.

One consistent finding across the two forcefields during the adsorption to methyl SAMs is motif availability. The motifs RGD and PHSRN were mostly unavailable, buried in the surface. It was found before that osteoblast-like cells on the methyl SAMs coated with $FnIII_{7\text{-}10}$ do not form actin cytoskeleton or focal adhesion sites [153]. It is therefore possible that the adsorption orientation defines the interactions. Another potential explanation is a denaturation of the fibronectin fragment. A fusion of the type III domain hydrophobic core with the hydrophobic surface would destroy the known fibronectin epitope.

On the two surfaces where adsorption took place I discussed the differences across forcefields. In Chapters 3 and 5 I highlighted the contributions from the van der Waals and suggested hydrophobicity to be an important factor in the adsorption. It is interesting therefore to compare the adsorption across the EA and methyl SAMs when simulated with the forcefield C36m. It was found that the adsorption of the 10th domain is largely the same across the two. The same key residues are present on both surfaces. In other words, there is enough similarity between EA and Methyl SAMs for the 10th domain to adsorb in the same way, but also enough difference for the 9th domain to show a very different adsorption behaviour. The 9th domain adsorbs to Methyl SAMs often, but it almost never does to EA SAMs. This similarity in adsorption of the 10th domain also explains why the RGD motif is unavailable for binding on both EA and methyl SAMs.

Whereas the availability of the two motifs was consistent on the methyl SAMs across the forcefields, it was not the case for EA SAMs. I showed that the RGD and PHSRN motifs were largely available for binding when the two domains adsorbed to the EA SAMs when the C36

forcefield is used. I suggested this as an explanation for the biological activity observed during adsorption to poly(ethyl acrylate) polymers. However, when the C36m forcefield is used, the adsorption uses the opposite site of the protein (Chapter 4). This adsorption state made the motifs unavailable for potential integrin-binding. This inconsistency is not easy to resolve. It might be possible that the adsorption state observed with C36 is as valid as the one observed in C36m. A further investigation into the parametrisation of the surface is necessary.

Moreover, it is important to highlight the confounding factors across the simulations with different forcefields, despite the efforts taken to minimise them. One larger difference is that the simulations with C36 forcefield were carried out in the NVT ensemble with the air-liquid interface in the system, whereas simulations with C36m forcefield accounted for pressure with the NPT ensemble. Another confounding factor is the periodic boundary condition (PBC) and the approach taken to calculate the long range electrostatics interactions with Particle Mesh Ewald (PME). With C36, due to the use of the gold forcefield GolP-CHARMM [141] the periodic boundary condition was created in the *xy* dimension, and therefore PME applied force and potential corrections in the *z* dimension. With C36m, PBC was applied in *xyz* dimensions, removing the need for any corrections. Ultimately, however, a more direct comparison to experimental data need to take place to resolve these dilemmas.

Orientation During the adsorption to methyl SAMs with the forcefield C36 the interdomain orientation of $FnIII_{9\text{-}10}$ changed. This was first noticed when the two motifs on the same side of the protein, RGD and PHSRN, diverged in their binding availability during adsorption. This change in interdomain orientation appeared to be stabilised by two hydrogen bonds between the two residues Arg1493 (from the RGD motif) and Asp1334 on the 9^{th} domain.

In order to understand the interdomain orientations, I simulated the two domains in bulk water. This led to the unexpected observation that the 9^{th} domain lost its tertiary structure. A β - strand detached from the hydrophobic core (Chapter 6). However, it was expected that the domain should remain stable at this temperature. I hypothesized that the interactions between the two domains was too strong which could have caused this structure unfolding. The application of the latest version of the forcefield, C36m, fixed this problem. C36m was

refined to help with the representation of intrinsically disordered proteins which was achieved by modifying the corrective CMAP potential of the backbone. It is possible therefore that any other simulations of similarly modular proteins could be affected by the overestimated protein-protein interactions.

The analysis of the interdomain orientation preferences of $FnIII_{9\text{-}10}$ was carried out with the forcefield C36m. Altogether 24 different systems with different interdomain orientations were simulated to understand if the two domains have any preference for particular conformations. The results showed one major interdomain orientation which also is the orientation found in the crystal structure (PDB:1FNF). However, two other orientations were found. This variety in the interdomain orientations is in agreement with the NMR structures of the two domains [155]. However, the ability of the two domains to assume different interdomain orientation has potential implications for the common view that the distance between the RGD and PHSRN motif is particularly important for the integrin adsorption. A mutant $FnIII_{9\text{-}10}$ with an extended linker between the two domains adsorbs poorly to the integrin $\alpha 3\beta 1$ in comparison to another mutant that stabilises the interdomain orientation [162]. Further, previous simulations have highlighted the importance of the 32 Å distance between the two motifs, suggesting an intermediate state with the distance 55 Å that behaves like a mechanical switch [152].

In one of the smaller cluster, $FnIII_{9\text{-}10}$ domains had a significant bend. In this cluster the angle between the two domains is around 60°. This is a significant departure from the main cluster where the structure fluctuates around the interdomain angle of 180°. In this smaller cluster the RGD and PHSRN motifs are particularly close to each other. This inclination to bend in one direction might be of importance to the formation of the compact fibronectin conformation.

9.1 Future Work

The approach taken to clustering the adsorption states in Chapter 8 can be further expanded. With well defined adsorption states, the transition rate can be calculated, which in this work was tracked manually. The transition rate in turn has the potential to be used in the discussion

of sampling and convergence. One might notice parallels with Markov State Modelling, one of the established tools in the molecular dynamics community [163].

While clustering the adsorption state I used the shortest distances from each residue to the surface. Further approaches to defining adsorption states could be explored. Ideally, additional variables such as hydrogen bonds would be taken into consideration in the process. Once the adsorption has been well characterised, active forces could be employed in order to understand the adsorption energies. Pulling biomolecules from the surface could be more directly comparable to experimental techniques such as atomic force microscopy.

Besides distances to the surface and adsorption, other quantities can help us understand the nature of adsorption. These should be more systematically explored in concert with the approaches used in this thesis. For example, using displacement and diffusion can help capture the more elusive surface-protein interactions [89].

In addition to adsorption, understanding the interdomain conformations between the different fibronectin domain pairs could help us understand how the fibronectin domains "fold" together to form the compact quaternary shape.

Fibronectin adsorb to many SAMs and have been observed to create fibronectin networks on EA and methyl SAMs. However, these networks are not equivalent and affects cells in different ways. Coarse-graining fibronectin to understand the different topologies the molecule can create could help with understanding the fibronectin networks, their physical properties, or their ability to expose important binding sites.

Ultimately, one of the keys to understanding fibronectin fibrillogenesis is to refine out understanding of how fibronectin interacts with the integrin receptors. Simulating such interactions could help describe the way in which integrins use the RGD and PHSRN motifs to initiate fibrillogenesis, and how that is linked to interdomain orientation.

9.2 The Future of Molecular Dynamics

Adsorption is a complex phenomena that depends on many factors. Molecular dynamics, even with its limitations, offers atomistic detail of the interactions between biomolecules and surfaces. Here, I would like to offer my vision of this technology and its impact on the field in the near future.

Moore's law currently defined as the doubling of the number of transistors on a chip every two years is being redefined. It is now used to describe the exponential decrease in computational cost. I experienced this first-hand during this PhD. Fast network infiniband-connected computers or nodes are very expensive. The Xeon CPUs, which are famously expensive, cost as much as the networking. However, the advent of GPUs is changing computational research. A single GPU such as Nvidia Volta 100 can be as fast as multiple networked Xeon-nodes. For this reason one molecular dynamics package, OpenMM, focused on supporting only GPUs, without providing support for multi-node computation [164, 138].

Another big promise is presented by the Anton supercomputer [165]. Despite relying on old technology, it can generate almost 60 μs per day even for a system including up to 100 thousand particles. Obtaining such a long simulation takes months on the best commercially available hardware. This specialised-hardware approach is complemented by parallel developments in another technology: field-programmable gate arrays (FPGA). I believe that in the near future simulations will use hardware such as FPGAs. One sign of this possibility is the newly released FPGA by the Intel corporation which supports native floating-point operations, OpenCL to make them more user friendly. This is while offering as much raw computing power as the best GPUs on the market.

The computational resources will be only as useful as the forcefields. The software on this front has evolved rapidly too. For example, open source software has been released to ease parametrising new forcefields [129]. This new promising approach showed that the water model TIP3P can be significantly refined, thanks to the well designed software and the available data. Groups including D.E. Shaw Research (the group behind the creation of Anton) put significant effort into refining the existing forcefields [166].

During my PhD I witnessed a proliferation of software in the field of molecular dynamics. Furthermore, I had the pleasure to contribute to two of them: MDAnalysis and PyMOL. Well written software improves usability, user productivity, quality, speed, flexibility, and ease which which it can be further refined by the community. Therefore, that the rapid evolution of both hardware and software will continue to be two important factors in the growth of molecular dynamics.

9.3 Limitations of my Methodology

The limitations of my methodology have been largely covered in the discussions. However, there is a one more limitation that I would like to highlight here.

In simulations I used only 1 or 2 domains from fibronectin which has 29 to 31 of them. A system containing a pair of solvated domains is sufficiently large that good hardware has to be used for many days (or weeks) to obtain enough data. Moreover, adding a surface to the system increases its size, which impacts negatively on the simulation performance. Therefore, in order to obtain meaningful simulation time, the use of larger protein fragments is not feasible. Even with access to the best computing resources, the molecular dynamics systems is going to remain an important constraint. For this reason dividing "large" problems into smaller manageable parts will remain a crucial approach in the field.

This brings us to extrapolating results from two domains to reveal the properties of fibronectin. During this thesis, in order to interpret the outcomes, I compared my results with the available experimental data. In some cases, with access to an NMR structure, a more exact comparison can take place. However, in the field of biology, most of the experimental data is qualitative. Furthermore, the vision is for the simulations to provide more information than there is in the existing publications.

In the near future the results will still have to be verified experimentally, which makes the simulations a complementary interpretation tool. However, as the predictive power of simulations increases, the intrinsic value of the simulations should gain more acceptance and more serious

consideration.

Bibliography

[1] D. J. Leahy, I. Aukhil, and H. P. Erickson, "2.0 Å Crystal Structure of a Four-Domain Segment of Human Fibronectin Encompassing the RGD Loop and Synergy Region," *Cell*, vol. 84, pp. 155–164, jan 1996.

[2] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.," *Biopolymers*, vol. 22, pp. 2577–637, dec 1983.

[3] R. Gerum, "pylustrator: Code generation for reproducible figures for publication," oct 2019.

[4] D. Banoriya, R. Purohit, and R. Dwivedi, "Advanced Application of Polymer based Biomaterials," *Materials Today: Proceedings*, vol. 4, pp. 3534–3541, jan 2017.

[5] M. Mir, M. N. Ali, A. Barakullah, A. Gulzar, M. Arshad, S. Fatima, and M. Asad, "Synthetic polymeric biomaterials for wound healing: a review," *Progress in Biomaterials*, vol. 7, pp. 1–21, mar 2018.

[6] P. Roach, D. Farrar, and C. C. Perry, "Interpretation of Protein Adsorption: Surface-Induced Conformational Changes," 2005.

[7] M. Geetha, A. Singh, R. Asokamani, and A. Gogia, "Ti based biomaterials, the ultimate choice for orthopaedic implants – A review," *Progress in Materials Science*, vol. 54, pp. 397–425, may 2009.

[8] M. B. Rahmany and M. Van Dyke, "Biomimetic approaches to modulate cellular adhesion in biomaterials: A review," *Acta Biomaterialia*, vol. 9, pp. 5431–5437, mar 2013.

[9] M. Saini, Y. Singh, P. Arora, V. Arora, and K. Jain, "Implant biomaterials: A comprehensive review.," *World journal of clinical cases*, vol. 3, pp. 52–7, jan 2015.

[10] L. Montanaro, P. Speziale, D. Campoccia, S. Ravaioli, I. Cangini, G. Pietrocola, S. Giannini, and C. R. Arciola, "Scenery of Staphylococcus implant infections in orthopedics," *Future Microbiology*, vol. 6, pp. 1329–1349, nov 2011.

[11] D. Campoccia, L. Montanaro, and C. R. Arciola, "A review of the biomaterials technologies for infection-resistant surfaces," *Biomaterials*, vol. 34, pp. 8533–8554, nov 2013.

[12] S. Franz, S. Rammelt, D. Scharnweber, and J. C. Simon, "Immune responses to implants – A review of the implications for the design of immunomodulatory biomaterials," *Biomaterials*, vol. 32, pp. 6692–6709, oct 2011.

[13] M. Rabe, D. Verdes, and S. Seeger, "Understanding protein adsorption phenomena at solid surfaces," *Advances in Colloid and Interface Science*, vol. 162, pp. 87–106, feb 2011.

[14] "The International Consortium of Investigative Journalists (ICIJ)." `https://www.icij.org/investigations/implant-files/`. Accessed: 2020-02-09.

[15] "The International Consortium of Investigative Journalists (ICIJ)." `https://web.archive.org/save/https://www.telegraph.co.uk/news/health/news/9631974/Faulty-medical-implants-investigation-Patients-failed-by-poor-implant-regulation-s.html`. Accessed: 2020-02-12.

[16] S. Kurtz, K. Ong, E. Lau, F. Mowat, and M. Halpern, "Projections of Primary and Revision Hip and Knee Arthroplasty in the United States from 2005 to 2030," *The Journal of Bone & Joint Surgery*, vol. 89, pp. 780–785, apr 2007.

[17] A. R. Araújo, D. Soares da Costa, S. Amorim, R. L. Reis, R. A. Pires, and I. Pashkuleva, "Surfaces Mimicking Glycosaminoglycans Trigger Different Response of Stem Cells via Distinct Fibronectin Adsorption and Reorganization," *ACS Applied Materials & Interfaces*, vol. 8, pp. 28428–28436, oct 2016.

[18] A. Domogatskaya, S. Rodin, A. Boutaud, and K. Tryggvason, "Laminin-511 but Not -332, -111, or -411 Enables Mouse Embryonic Stem Cell Self-Renewal In Vitro," *Stem Cells*, vol. 26, pp. 2800–2809, nov 2008.

[19] J. D. Andrade, V. Hlady, and A. P. Wei, "Adsorption of complex proteins at interfaces," *Pure and Applied Chemistry*, vol. 64, no. 11, pp. 1777–1781, 1992.

[20] E. Mavropoulos, A. M. Costa, L. T. Costa, C. A. Achete, A. Mello, J. M. Granjeiro, and A. M. Rossi, "Adsorption and bioactivity studies of albumin onto hydroxyapatite surface," *Colloids and Surfaces B: Biointerfaces*, vol. 83, pp. 1–9, mar 2011.

[21] K. Cai, J. Bossert, and K. D. Jandt, "Does the nanometre scale topography of titanium influence protein adsorption and cell proliferation?," *Colloids and Surfaces B: Biointerfaces*, vol. 49, pp. 136–144, may 2006.

[22] U. Hersel, C. Dahmen, and H. Kessler, "RGD modified polymers: biomaterials for stimulated cell adhesion and beyond," *Biomaterials*, vol. 24, pp. 4385–4415, nov 2003.

[23] E. A. dos Santos, M. Farina, G. A. Soares, and K. Anselme, "Surface energy of hydroxyapatite and β-tricalcium phosphate ceramics driving serum protein adsorption and osteoblast adhesion," *Journal of Materials Science: Materials in Medicine*, vol. 19, pp. 2307–2316, jun 2008.

[24] K. M. Yamada and K. Olden, "Fibronectins—adhesive glycoproteins of cell surface and blood," *Nature*, vol. 275, pp. 179–184, sep 1978.

[25] G. A. Di Lullo, S. M. Sweeney, J. Korkko, L. Ala-Kokko, and J. D. San Antonio, "Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen.," *The Journal of biological chemistry*, vol. 277, pp. 4223–31, feb 2002.

[26] S. Teoh, "Fatigue of biomaterials: a review," *International Journal of Fatigue*, vol. 22, pp. 825–837, nov 2000.

[27] F. Wu, D. D. W. Lin, J. H. Chang, C. Fischbach, L. A. Estroff, and D. Gourdon, "Effect of the Materials Properties of Hydroxyapatite Nanoparticles on Fibronectin Deposition and Conformation," *Crystal Growth & Design*, vol. 15, pp. 2452–2460, may 2015.

[28] K. Wang, C. Zhou, Y. Hong, and X. Zhang, "A review of protein adsorption on bioceramics," *Interface Focus*, vol. 2, pp. 259–277, jun 2012.

[29] M. Kaur and K. Singh, "Review on titanium and titanium based alloys as biomaterials for orthopaedic applications," *Materials Science and Engineering: C*, vol. 102, pp. 844–862, sep 2019.

[30] F. A. Shah, M. Trobos, P. Thomsen, and A. Palmquist, "Commercially pure titanium (cp-Ti) versus titanium alloy (Ti6Al4V) materials as bone anchored implants — Is one truly better than the other?," *Materials Science and Engineering: C*, vol. 62, pp. 960–966, may 2016.

[31] S. Ramakrishna, J. Mayer, E. Wintermantel, and K. W. Leong, "Biomedical applications of polymer-composite materials: a review," *Composites Science and Technology*, vol. 61, pp. 1189–1224, jul 2001.

[32] D. Hyslop, A. Abdelkader, A. Cox, and D. Fray, "Electrochemical synthesis of a biomedically important Co–Cr alloy," *Acta Materialia*, vol. 58, pp. 3124–3130, may 2010.

[33] S. Nag, R. Banerjee, and H. Fraser, "Microstructural evolution and strengthening mechanisms in Ti–Nb–Zr–Ta, Ti–Mo–Zr–Fe and Ti–15Mo biocompatible alloys," *Materials Science and Engineering: C*, vol. 25, pp. 357–362, may 2005.

[34] I. GOTMAN, "Characteristics of Metals Used in Implants," *Journal of Endourology*, vol. 11, pp. 383–389, dec 1997.

[35] K. L. Wapner, "Implications of metallic corrosion in total knee arthroplasty.," *Clinical orthopaedics and related research*, vol. 271, pp. 12–20, oct 1991.

[36] N. Kukreja, Y. Onuma, J. Daemen, and P. W. Serruys, "The future of drug-eluting stents," *Pharmacological Research*, vol. 57, pp. 171–180, mar 2008.

[37] M. B. Hovgaard, K. Rechendorff, J. Chevallier, M. Foss, and F. Besenbacher, "Fibronectin Adsorption on Tantalum: The Influence of Nanoroughness," *The Journal of Physical Chemistry B*, vol. 112, pp. 8241–8249, jul 2008.

[38] M. Lehnert, M. Gorbahn, C. Rosin, M. Klein, I. Koper, B. Al-Nawas, W. Knoll, and M. Veith, "Adsorption and Conformation Behavior of Biotinylated Fibronectin on Streptavidin-Modified TiO _X Surfaces Studied by SPR and AFM," *Langmuir*, vol. 27, pp. 7743–7751, jun 2011.

[39] J. W. Tamkun and R. O. Hynes, "Plasma fibronectin is synthesized and secreted by hepatocytes," *Journal of Biological Chemistry*, vol. 258, no. 7, pp. 4641–4647, 1983.

[40] G. Zerlauth and G. Wolf, "Plasma fibronectin as a marker for cancer and other diseases," *The American Journal of Medicine*, vol. 77, pp. 685–689, oct 1984.

[41] F. A. Moretti, A. K. Chauhan, A. Iaconcig, F. Porro, F. E. Baralle, and A. F. Muro, "A major fraction of fibronectin present in the extracellular matrix of tissues is plasma-derived.," *The Journal of biological chemistry*, vol. 282, pp. 28057–62, sep 2007.

[42] M. M. Martino and J. A. Hubbell, "The 12th–14th type III repeats of fibronectin function as a highly promiscuous growth factor-binding domain," *The FASEB Journal*, vol. 24, pp. 4711–4721, dec 2010.

[43] B. Geiger, A. Bershadsky, R. Pankov, and K. M. Yamada, "Transmembrane crosstalk between the extracellular matrix–cytoskeleton crosstalk.," *Nature reviews. Molecular cell biology*, vol. 2, pp. 793–805, nov 2001.

[44] R. Pankov, "Fibronectin at a glance," *Journal of Cell Science*, vol. 115, pp. 3861–3863, oct 2002.

[45] J. E. Schwarzbauer, "Identification of the fibronectin sequences required for assembly of a fibrillar matrix.," *The Journal of cell biology*, vol. 113, pp. 1463–73, jun 1991.

[46] E. L. George, E. N. Georges-Labouesse, R. S. Patel-King, H. Rayburn, and R. O. Hynes, "Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin.," *Development (Cambridge, England)*, vol. 119, pp. 1079–91, dec 1993.

[47] J. Sottile, D. Hocking, and K. Langenbach, "Fibronectin polymerization stimulates cell growth by RGD-dependent and -independent mechanisms," *J. Cell Sci.*, vol. 113, pp. 4287–4299, dec 2000.

[48] C. M. Williams, A. J. Engler, R. D. Slone, L. L. Galante, and J. E. Schwarzbauer, "Fibronectin expression modulates mammary epithelial cell proliferation during acinar differentiation.," *Cancer research*, vol. 68, pp. 3185–92, may 2008.

[49] H. Ohtsubo, T. Okada, K. Nozu, Y. Takaoka, A. Shono, K. Asanuma, L. Zhang, K. Nakanishi, M. Taniguchi-Ikeda, H. Kaito, K. Iijima, and S.-i. Nakamura, "Identification of mutations in FN1 leading to glomerulopathy with fibronectin deposits," *Pediatric Nephrology*, vol. 31, pp. 1459–1467, sep 2016.

[50] I. Ishimoto, E. Sohara, E. Ito, T. Okado, T. Rai, and S. Uchida, "Fibronectin glomerulopathy," *Clinical Kidney Journal*, vol. 6, pp. 513–515, oct 2013.

[51] I. Brčić, L. Brčić, D. Kuzmanić, M. Ćorić, and M. Ćorić, "Fibronectin Glomerulopathy in a 34-year-old Man: A Case Report," *Ultrastructural Pathology*, vol. 34, pp. 240–242, aug 2010.

[52] M. Nadamuni, R. Piras, S. Mazbar, J. P. Higgins, and N. Kambham, "Fibronectin Glomerulopathy: An Unusual Cause of Adult-Onset Nephrotic Syndrome," *American Journal of Kidney Diseases*, vol. 60, pp. 839–842, nov 2012.

[53] Z. Han and Lu, "Targeting fibronectin for cancer imaging and therapy," *J. Mater. Chem. B*, vol. 32, pp. 303–315, 2017.

[54] M. W. Pickup, J. K. Mouw, and V. M. Weaver, "The extracellular matrix modulates the hallmarks of cancer," *EMBO reports*, vol. 15, pp. 1243–1253, dec 2014.

[55] S. A. An Seong, M. Llinás, J. Jimenez-Barbero, and T. E. Petersen, "The Two Polypeptide Chains in Fibronectin Are Joined in Antiparallel Fashion:NMR Structural Characterization," *Biochemistry*, vol. 31, no. 41, pp. 9927–9933, 1992.

[56] E. White, F. Baralle, and A. Muro, "New insights into form and function of fibronectin splice variants," *The Journal of Pathology*, vol. 216, pp. 1–14, sep 2008.

[57] M. Rocco, E. Infusini, M. G. Daga, L. Gogioso, and C. Cuniberti, "Models of fibronectin.," *The EMBO journal*, vol. 6, pp. 2343–9, aug 1987.

[58] K. J. Johnson, H. Sage, G. Briscoe, and H. P. Erickson, "The Compact Conformation of Fibronectin Is Determined by Intramolecular Ionic Interactions," *Journal of Biological Chemistry*, vol. 274, pp. 15473–15479, may 1999.

[59] M. Y. Khan, M. S. Medow, and S. A. Newman, "Unfolding transitions of fibronectin and its domains. Stabilization and structural alteration of the N-terminal domain by heparin.," *The Biochemical journal*, vol. 270, pp. 33–8, aug 1990.

[60] V. P. Ivanova, "Fibronectins: Structural-functional relationships," *Journal of Evolutionary Biochemistry and Physiology*, vol. 53, pp. 450–464, nov 2017.

[61] P. Singh, C. Carraher, and J. E. Schwarzbauer, "Assembly of fibronectin extracellular matrix.," *Annual review of cell and developmental biology*, vol. 26, pp. 397–419, jan 2010.

[62] L. M. Maurer, W. Ma, and D. F. Mosher, "Dynamic structure of plasma fibronectin," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 51, pp. 213–227, jul 2016.

[63] E. Österlund, "The secondary structure of human plasma fibronectin: conformational changes induced by acidic pH and elevated temperatures; a circular dichroic study," *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, vol. 955, pp. 330–336, aug 1988.

[64] J. R. Potts and I. D. Campbell, "Structure and function of fibronectin modules," *Matrix Biology*, vol. 15, pp. 313–320, nov 1996.

[65] S. E. D'Souza, M. H. Ginsberg, and E. F. Plow, "Arginyl-glycyl-aspartic acid (RGD): a cell adhesion motif," *Trends in Biochemical Sciences*, vol. 16, pp. 246–250, jan 1991.

[66] S. Aota, M. Nomizu, and K. M. Yamada, "The short amino acid sequence Pro-His-Ser-Arg-Asn in human fibronectin enhances cell-adhesive function.," *The Journal of biological chemistry*, vol. 269, pp. 24756–61, oct 1994.

[67] S. D. Redick, D. L. Settles, G. Briscoe, and H. P. Erickson, "Defining fibronectin's cell adhesion synergy site by site-directed mutagenesis.," *The Journal of cell biology*, vol. 149, pp. 521–7, apr 2000.

[68] J. L. Sechler, Y. Takada, and J. E. Schwarzbauer, "Altered rate of fibronectin matrix assembly by deletion of the first type III repeats.," *The Journal of cell biology*, vol. 134, pp. 573–83, jul 1996.

[69] C. Wu, V. M. Keivens, T. E. O'Toole, J. A. McDonald, and M. H. Ginsberg, "Integrin activation and cytoskeletal interaction are essential for the assembly of a fibronectin matrix.," *Cell*, vol. 83, pp. 715–24, dec 1995.

[70] C. Zhong, M. Chrzanowska-Wodnicka, J. Brown, A. Shaub, A. M. Belkin, and K. Burridge, "Rho-mediated Contractility Exposes a Cryptic Site in Fibronectin and Induces Fibronectin Matrix Assembly," *The Journal of Cell Biology*, vol. 141, no. 2, 1998.

[71] S. Miyamoto, S. K. Akiyama, and K. M. Yamada, "Synergistic roles for receptor occupancy and aggregation in integrin transmembrane function.," *Science (New York, N.Y.)*, vol. 267, pp. 883–5, feb 1995.

[72] M. Cantini, Critina Gonazlez-Garcia, V. Llopis-Hernandez, and M. Salmerón-Sánchez, "Material-Driven Fibronectin Fibrillogenesis," 2012.

[73] V. Vogel, "Fibronectin in a Surface-Adsorbed State," pp. 505–518, may 1995.

[74] V. Vogel, "Unraveling the Mechanobiology of Extracellular Matrix," *Annual Review of Physiology*, vol. 80, pp. 353–387, feb 2018.

[75] D. C. Hocking, J. Sottile, and P. J. McKeown-Longo, "Fibronectin's III-1 module contains a conformation-dependent binding site for the amino-terminal region of fibronectin.," *The Journal of biological chemistry*, vol. 269, pp. 19183–7, jul 1994.

[76] K. C. Ingham, S. A. Brew, S. Huff, and S. V. Litvinovich, "Cryptic self-association sites in type III modules of fibronectin.," *The Journal of biological chemistry*, vol. 272, pp. 1718–24, jan 1997.

[77] A. G. Hemmersam, K. Rechendorff, M. Foss, D. S. Sutherland, and F. Besenbacher, "Fibronectin adsorption on gold, Ti-, and Ta-oxide investigated by QCM-D and RSA modelling," *Journal of Colloid and Interface Science*, vol. 320, pp. 110–116, apr 2008.

[78] M. Raoufi, M. J. Hajipour, S. M. Kamali Shahri, I. Schön, U. Linne, and M. Mahmoudi, "Probing Fibronectin Conformation on Protein Corona Layer around Nanoparticles," *Nanoscale*, 2017.

[79] S. R. Sousa, M. M. Brás, P. Moradas-Ferreira, and M. A. Barbosa†, "Dynamics of Fibronectin Adsorption on TiO2 Surfaces," 2007.

[80] L. Lv, K. Li, Y. Xie, Y. Cao, and X. Zheng, "Enhanced osteogenic activity of anatase TiO2 film: Surface hydroxyl groups induce conformational changes in fibronectin," *Materials Science and Engineering: C*, vol. 78, pp. 96–104, 2017.

[81] D. H. K. Nguyen, V. T. H. Pham, M. Al Kobaisi, C. Bhadra, A. Orlowska, S. Ghanaati, B. M. Manzi, V. A. Baulin, S. Joudkazis, P. Kingshott, R. J. Crawford, and E. P. Ivanova, "Adsorption of Human Plasma Albumin and Fibronectin onto Nanostructured Black Silicon Surfaces," *Langmuir*, vol. 32, pp. 10744–10751, oct 2016.

[82] P. J. Molino, M. J. Higgins, P. C. Innis, R. M. I. Kapsa, and G. G. Wallace, "Fibronectin and Bovine Serum Albumin Adsorption and Conformational Dynamics on Inherently Conducting Polymers: A QCM-D Study," *Langmuir*, vol. 28, pp. 8433–8445, jun 2012.

[83] N. Giamblanco, G. Zhavnerko, N. Tuccitto, A. Licciardello, and G. Marletta, "Coadsorption-dependent orientation of fibronectin epitopes at hydrophilic gold surfaces," *Soft Matter*, vol. 8, p. 8370, jul 2012.

[84] C. R. Wittmer and P. R. Van Tassel, "Probing adsorbed fibronectin layer structure by kinetic analysis of monoclonal antibody binding," *Colloids and Surfaces B: Biointerfaces*, vol. 41, pp. 103–109, mar 2005.

[85] B. G. Keselowsky, D. M. Collard, and A. J. García, "Surface chemistry modulates fibronectin conformation and directs integrin binding and specificity to control cell adhesion," *Journal of Biomedical Materials Research Part A*, vol. 66A, pp. 247–259, aug 2003.

[86] P. Y. Meadows and G. C. Walker, "Force Microscopy Studies of Fibronectin Adsorption and Subsequent Cellular Adhesion to Substrates with Well-Defined Surface Chemistries," 2005.

[87] W. Norde, T. A. Horbett, and J. L. Brash, "Proteins at Interfaces III: Introductory Overview," pp. 1–34, jan 2012.

[88] K. Kubiak-Ossowska and P. A. Mulheran, "What Governs Protein Adsorption and Immobilization at a Charged Solid Surface?," *Langmuir*, vol. 26, pp. 7690–7694, jun 2010.

[89] K. Kubiak-Ossowska, B. Jachimska, and P. A. Mulheran, "How Negatively Charged Proteins Adsorb to Negatively Charged Surfaces: A Molecular Dynamics Study of BSA Adsorption on Silica," *The Journal of Physical Chemistry B*, vol. 120, pp. 10463–10468, oct 2016.

[90] C. A. Haynes and W. Norde, "Globular proteins at solid/liquid interfaces," *Colloids and Surfaces B: Biointerfaces*, vol. 2, pp. 517–566, jul 1994.

[91] R. C. Bernardi, M. C. Melo, and K. Schulten, "Enhanced sampling techniques in molecular dynamics simulations of biological systems," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1850, pp. 872–877, may 2015.

[92] X. Wang, Z. Li, H. Li, S. Ruan, and J. Gu, "Influence of nanoscale surface curvature of rutile on fibronectin adsorption by atomistic simulations," *Journal of Materials Science*, vol. 52, pp. 13512–13521, dec 2017.

[93] E. Liamas, K. Kubiak-Ossowska, R. Black, O. Thomas, Z. Zhang, P. Mulheran, E. Liamas, K. Kubiak-Ossowska, R. A. Black, O. R. Thomas, Z. J. Zhang, and P. A. Mulheran, "Adsorption of Fibronectin Fragment on Surfaces Using Fully Atomistic Molecular Dynamics Simulations," *21*, vol. 19, p. 3321, oct 2018.

[94] K. Kubiak-Ossowska, P. A. Mulheran, and W. Nowak, "Fibronectin module FN(III)9 adsorption at contrasting solid model surfaces studied by atomistic molecular dynamics.," *The journal of physical chemistry. B*, vol. 118, pp. 9900–8, aug 2014.

[95] T. Li, L. Hao, J. Li, C. Du, and Y. Wang, "The Role of Ninth Type-III Domain of Fibronectin in the Mediation of Cell-binding Domain Adsorption on Surfaces with Different Chemistry," *Langmuir*, p. acs.langmuir.8b01937, jul 2018.

[96] C. Liao, Y. Xie, and J. Zhou, "Computer simulations of fibronectin adsorption on hydroxyapatite surfaces," *RSC Advances*, vol. 4, no. 30, p. 15759, 2014.

[97] D. Gugutkov, C. González-García, J. C. Rodríguez Hernández, G. Altankov, and M. Salmerón-Sánchez, "Biological activity of the substrate-induced fibronectin network: insight into the third dimension through electrospun fibers.," *Langmuir : the ACS journal of surfaces and colloids*, vol. 25, pp. 10893–900, sep 2009.

[98] M. Salmerón-Sánchez, P. Rico, D. Moratal, T. T. Lee, J. E. Schwarzbauer, and A. J. García, "Role of material-driven fibronectin fibrillogenesis in cell differentiation.," *Biomaterials*, vol. 32, pp. 2099–105, mar 2011.

[99] R. Emch, F. Zenhausern, M. Jobin, M. Taborelli, and P. Descouts, "Morphological difference between fibronectin sprayed on mica and on PMMA.," *Ultramicroscopy*, pp. 1155–60, jul 1992.

[100] F. Bathawab, M. Bennett, M. Cantini, J. Reboud, M. J. Dalby, and M. Salmerón-Sánchez, "Lateral Chain Length in Polyalkyl Acrylates Determines the Mobility of Fibronectin at the Cell/Material Interface.," *Langmuir : the ACS journal of surfaces and colloids*, jan 2016.

[101] J. Ballester-Beltrán, P. Rico, D. Moratal, W. Song, J. F. Mano, and M. Salmerón-Sánchez, "Role of superhydrophobicity in the biological activity of fibronectin at the cell–material interface," *Soft Matter*, vol. 7, p. 10803, nov 2011.

[102] C. Ribeiro, J. A. Panadero, V. Sencadas, S. Lanceros-Méndez, M. N. Tamaño, D. Moratal, M. Salmerón-Sánchez, and J. L. Gómez Ribelles, "Fibronectin adsorption and cell response on electroactive poly(vinylidene fluoride) films," *Biomedical Materials*, vol. 7, p. 035004, jun 2012.

[103] M. Bergkvist, J. Carlsson, and S. Oscarsson, "Surface-dependent conformations of human plasma fibronectin adsorbed to silica, mica, and hydrophobic surfaces, studied with use of Atomic Force Microscopy," *Journal of Biomedical Materials Research*, vol. 64A, pp. 349–356, feb 2003.

[104] L. A. Culp and C. N. Sukenik, "Cell type-specific modulation of fibronectin adhesion functions on chemically-derivatized self-assembled monolayers," *Journal of Biomaterials Science, Polymer Edition*, vol. 9, pp. 1161–1176, jan 1998.

[105] R. Agarwal, C. González-García, B. Torstrick, R. E. Guldberg, M. Salmerón-Sánchez, and A. J. García, "Simple coating with fibronectin fragment enhances stainless steel screw osseointegration in healthy and osteoporotic rats," *Biomaterials*, vol. 63, pp. 137–145, sep 2015.

[106] V. Llopis-Hernandez, M. Cantini, C. Gonzalez-Garcia, Z. A. Cheng, J. Yang, P. M. Tsimbouri, A. J. Garcia, M. J. Dalby, and M. Salmeron-Sanchez, "Material-driven fibronectin assembly for high-efficiency presentation of growth factors," *Science Advances*, vol. 2, pp. e1600188–e1600188, aug 2016.

[107] J. D. WATSON and F. H. C. CRICK, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature*, vol. 171, pp. 737–738, apr 1953.

[108] B. J. Alder and T. E. Wainwright, "Studies in Molecular Dynamics. I. General Method," *The Journal of Chemical Physics*, vol. 31, pp. 459–466, aug 1959.

[109] P. Sneha and C. George Priya Doss, "Molecular Dynamics," in *Advances in protein chemistry and structural biology*, vol. 102, pp. 181–224, 2016.

[110] L. Hovan, V. Oleinikovas, H. Yalinca, A. Kryshtafovych, G. Saladino, and F. L. Gervasio, "Assessment of the model refinement category in CASP12," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 152–167, mar 2018.

[111] J. Henriques, C. Cragnell, and M. Skepö, "Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment," *Journal of Chemical Theory and Computation*, vol. 11, pp. 3420–3431, jul 2015.

[112] F. Khalili-Araghi, J. Gumbart, P.-C. Wen, M. Sotomayor, E. Tajkhorshid, and K. Schulten, "Molecular dynamics simulations of membrane channels and transporters.," *Current opinion in structural biology*, vol. 19, pp. 128–37, apr 2009.

[113] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, jan 2000.

[114] W. F. van Gunsteren and H. J. C. Berendsen, "Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry," *Angewandte Chemie International Edition in English*, vol. 29, pp. 992–1023, sep 1990.

[115] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt, and H. B. Yu, "Biomolecular modeling: Goals, problems, perspectives.," *Angewandte Chemie (International ed. in English)*, vol. 45, pp. 4064–92, jun 2006.

[116] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera,

D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins.," *The journal of physical chemistry. B*, vol. 102, pp. 3586–616, apr 1998.

[117] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, "Definition and testing of the GROMOS force-field versions 54A7 and 54B7.," *European biophysics journal : EBJ*, vol. 40, pp. 843–56, jul 2011.

[118] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *Journal of the American Chemical Society*, vol. 117, pp. 5179–5197, may 1995.

[119] W. L. Jorgensen and J. Tirado-Rives, "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin," *Journal of the American Chemical Society*, vol. 110, pp. 1657–1666, mar 1988.

[120] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, "CHARMM36m: an improved force field for folded and intrinsically disordered proteins," *Nature Methods*, vol. 14, pp. 71–73, nov 2016.

[121] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, pp. 3684–3690, oct 1984.

[122] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *The Journal of Chemical Physics*, vol. 126, p. 014101, jan 2007.

[123] W. G. Hoover and B. L. Holian, "Kinetic moments method for the canonical ensemble distribution," *Physics Letters A*, vol. 211, pp. 253–257, feb 1996.

[124] H. A. Posch, W. G. Hoover, and F. J. Vesely, "Canonical dynamics of the Nosé oscillator: Stability, order, and chaos," *Physical Review A*, vol. 33, pp. 4253–4265, jun 1986.

[125] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied Physics*, vol. 52, pp. 7182–7190, dec 1981.

[126] S. Nosé and M. Klein, "Constant pressure molecular dynamics for molecular systems," *Molecular Physics*, vol. 50, pp. 1055–1076, dec 1983.

[127] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of Computational Physics*, vol. 23, pp. 327–341, mar 1977.

[128] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," *Journal of Computational Chemistry*, vol. 18, pp. 1463–1472, sep 1997.

[129] L.-P. Wang, T. J. Martinez, and V. S. Pande, "Building Force Fields: An Automatic, Systematic, and Reproducible Approach," *The Journal of Physical Chemistry Letters*, vol. 5, pp. 1885–1891, jun 2014.

[130] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, no. 2, p. 926, 1983.

[131] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. Mackerell, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles.," *Journal of chemical theory and computation*, vol. 8, pp. 3257–3273, sep 2012.

[132] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, "Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States," *The Journal of Physical Chemistry B*, vol. 119, pp. 5113–5123, apr 2015.

[133] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "MDAnalysis: a toolkit for the analysis of molecular dynamics simulations.," *Journal of computational chemistry*, vol. 32, pp. 2319–27, jul 2011.

[134] R. Gowers, M. Linke, J. Barnoud, T. Reddy, M. Melo, S. Seyler, J. Domański, D. Dotson, S. Buchoux, I. Kenney, and O. Beckstein, "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," in *Proceedings of the 15th Python in Science Conference*, pp. 98–105, 2016.

[135] M. Ester, M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," pp. 226—-231, 1996.

[136] L. C. Pardo, J. L. Tamarit, N. Veglio, F. J. Bermejo, and G. J. Cuello, "Comparison of short-range-order in liquid- and rotator-phase states of a simple molecular liquid: A reverse Monte Carlo and molecular dynamics analysis of neutron diffraction data," *Physical Review B*, vol. 76, p. 134203, oct 2007.

[137] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, jul 2015.

[138] C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B. L. de Groot, and H. Grubmüller, "Best bang for your buck: GPU nodes for GROMACS biomolecular simulations," *Journal of Computational Chemistry*, vol. 36, pp. 1990–2008, oct 2015.

[139] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform," *Journal of Cheminformatics*, vol. 4, p. 17, aug 2012.

[140] K. Vanommeslaeghe, E. P. Raman, A. D. MacKerell, and Jr, "Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges.," *Journal of chemical information and modeling*, vol. 52, pp. 3155–68, dec 2012.

[141] L. B. Wright, P. M. Rodger, S. Corni, and T. R. Walsh, "GolP-CHARMM: First-Principles Based Force Fields for the Interaction of Proteins with Au(111) and Au(100).," *Journal of chemical theory and computation*, vol. 9, pp. 1616–30, mar 2013.

[142] Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domański, David L. Dotson, Sébastien Buchoux, Ian M. Kenney, and Oliver Beckstein, "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," in *Proceedings of the 15th Python in Science Conference* (Sebastian Benthall and Scott Rostrup, eds.), pp. 98 – 105, 2016.

[143] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker, "PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations," *Nucleic Acids Research*, vol. 32, pp. W665–W667, jul 2004.

[144] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: application to microtubules and the ribosome.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 10037–41, aug 2001.

[145] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[146] J. Stone, "*An Efficient Library for Parallel Ray Tracing and Animation*," Master's thesis, Computer Science Department, University of Missouri-Rolla, April 1998.

[147] Schrödinger, LLC, "The PyMOL molecular graphics system, version 1.8." November 2015.

[148] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[149] R. Kumari, R. Kumar, A. Lynn, and A. Lynn, "g_mmpbsa —A GROMACS Tool for High-Throughput MM-PBSA Calculations," *Journal of Chemical Information and Modeling*, vol. 54, pp. 1951–1962, jul 2014.

[150] A. F. Oberhauser, C. Badilla-Fernandez, M. Carrion-Vazquez, and J. M. Fernandez, "The Mechanical Hierarchies of Fibronectin Observed with Single-molecule AFM," *Journal of Molecular Biology*, vol. 319, pp. 433–447, may 2002.

[151] D. Craig, A. Krammer, K. Schulten, and V. Vogel, "Comparison of the early stages of forced unfolding for fibronectin type III modules.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 5590–5, may 2001.

[152] A. Krammer, D. Craig, W. E. Thomas, K. Schulten, and V. Vogel, "A structural model for force regulated integrin binding to fibronectin's RGD-synergy site," *Matrix Biology*, vol. 21, no. 2, pp. 139–147, 2002.

[153] P. Rico, C. González-García, T. A. Petrie, A. J. García, and M. Salmerón-Sánchez, "Molecular assembly and biological activity of a recombinant fragment of fibronectin (FNIII7–10) on poly(ethyl acrylate)," *Colloids and Surfaces B: Biointerfaces*, vol. 78, pp. 310–316, jul 2010.

[154] S. V. Litvinovich and K. C. Ingham, "Interactions Between Type III Domains in the 110 kDa Cell-binding Fragment of Fibronectin," *Journal of Molecular Biology*, vol. 248, pp. 611–626, may 1995.

[155] V. Copié, Y. Tomita, S. K. Akiyama, S.-i. Aota, K. M. Yamada, R. M. Venable, R. W. Pastor, S. Krueger, and D. A. Torchia, "Solution structure and dynamics of linked cell attachment modules of mouse fibronectin containing the RGD and synergy regions: comparison with the human fibronectin crystal structure," *Journal of Molecular Biology*, vol. 277, pp. 663–682, apr 1998.

[156] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[157] K. T. Debiec, M. J. Whitley, L. M. I. Koharudin, L. T. Chong, and A. M. Gronenborn, "Integrating NMR, SAXS, and Atomistic Simulations: Structure and Dynamics of a Two-Domain Protein.," *Biophysical journal*, vol. 114, pp. 839–855, feb 2018.

[158] A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics," *WIREs Computational Molecular Science*, vol. 1, pp. 826–843, sep 2011.

[159] C. Kobayashi, Y. Matsunaga, R. Koike, M. Ota, and Y. Sugita, "Domain Motion Enhanced (DoME) Model for Efficient Conformational Sampling of Multidomain Proteins," *The Journal of Physical Chemistry B*, vol. 119, pp. 14584–14593, nov 2015.

[160] K. Moritsugu and J. C. Smith, "Coarse-Grained Biomolecular Simulation with REACH: Realistic Extension Algorithm via Covariance Hessian," *Biophysical Journal*, vol. 93, pp. 3460–3469, nov 2007.

[161] M. Ozboyaci, D. B. Kokh, and R. C. Wade, "Three steps to gold: mechanism of protein adsorption revealed by Brownian and molecular dynamics simulations," *Physical Chemistry Chemical Physics*, vol. 18, pp. 10191–10200, apr 2016.

[162] A. C. Brown, M. M. Dysart, K. C. Clarke, S. E. Stabenfeldt, and T. H. Barker, "Integrin $\alpha 3\beta 1$ Binding to Fibronectin Is Dependent on the Ninth Type III Repeat.," *The Journal of biological chemistry*, vol. 290, pp. 25534–47, oct 2015.

[163] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models," *Journal of Chemical Theory and Computation*, vol. 11, pp. 5525–5542, nov 2015.

[164] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," *PLOS Computational Biology*, vol. 13, p. e1005659, jul 2017.

[165] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L.-S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y.-H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young, "Anton 2: Raising the

Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer," in *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 41–53, IEEE, nov 2014.

[166] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states.," *Proceedings of the National Academy of Sciences of the United States of America*, p. 201800690, may 2018.

[167] M. K. Bieniek, V. Llopis-Hernandez, K. Douglas, M. Salmeron-Sanchez, and C. D. Lorenz, "Minor Chemistry Changes Alter Surface Hydration to Control Fibronectin Adsorption and Assembly into Nanofibrils," *Advanced Theory and Simulations*, p. 1900169, oct 2019.

Appendix A

Software Development

This Appendix focuses on software development I carried out during my PhD. Along with the attached code, a brief description and comments are provided. The first section "Analysis" contains several scripts which have been used to obtain and plot the results in the thesis. The scripts that have been used during the work on my publication [167] have been further uploaded to the github repository and made available under the creative commons license (`https://github.com/bieniekmateusz/publications/tree/master/Minor_Chemistry_Changes_Alter_Surface_Hydration_to_Control_Fibronectin_Adsorption_and_Assembly_into_Nanofibrils`). The last two sections focus on two open source contributions which have been made in collaboration with Paul Smith, King's College London. The first open source contribution was made to the python analysis package MDAnalysis, and in this Appendix further comments and description of the changes are provided. The second contribution was made to the molecular visualiser PyMOL, and it was carried out as part of the joint Warren L. DeLano Memorial PyMOL Open-Source Fellowship awarded by the Schrödinger company that maintains the PyMOL software.

A.1 Analysis

In this section the analysis scripts are placed along with the comments and description for any future potential users. The first script calculates the residue-residue distances over time which

are then saved to a file. This script is further complemented with maptlotlib code which plots the outcomes in the form of a heatmap. The second analysis script describes the hydration around the selected functional groups. However, the nature of the script is such that it can be more universally applied to other molecules to understand their positions with respect to each other over time. Similarly, this script is accompanied with a tutorial-like description. The third analysis concerns clustering which was used in the later parts of the thesis, and for this the provided code shows how to retrieve, compare, cluster and visualise the data from the trajectories.

Residue-Surface Distances

The first python script uses MDAnalysis to calculate the minimum distance between the heavy atoms of each residue and the heavy atoms in the surface (Self-assembled Monolayers). A small optimisation is introduced where the selections of heavy atoms for each residue are cached. This approach can be complemented with a parallel component. However, in order to properly optimise the performance for the parallel version of the code, a non-trivial approach is needed where the memory access and caching is considered carefully. Dividing the simulation into n chunks for n CPUs would likely be easy and fast, as long as solid state drives (SSD) are used.

```
1  #!/usr/bin/env python
2  # -*- coding: utf-8 -*-
3  """
4  Calculate the minimum distance from the heavy atoms of
5  each residue and the surface. Save the results in a file
6  with a tabular format with the following format:
7  # time(ps) res1aaa.res1id.res1index res2aaa.res2id.res2index
8
9  Further details can be found on the github link:
   ↪ https://github.com/bieniekmateusz/publications/
10
```

```
11  Please cite the following publication if you find this script useful:
12  Bieniek, M. K. et al. (2019) 'Minor Chemistry Changes Alter Surface
    ↪ Hydration to Control Fibronectin Adsorption and Assembly into
    ↪ Nanofibrils', Advanced Theory and Simulations. John Wiley & Sons, Ltd,
    ↪ p. 1900169. doi: 10.1002/adts.201900169.
13  """
14
15  import MDAnalysis
16  from MDAnalysis.analysis.distances import distance_array
17  import numpy as np
18
19  # open the trajectory with MDAnalysis
20  u = MDAnalysis.Universe('500ns_protCent_pbcMol.gro',
21  '500ns_centCA_pbcMol_step100ps.xtc')
22  # select the atoms in the substrate that are found on the surface
23  substrate = u.select_atoms("resname EA and (name C11 O1 O2 C12 C13)")
24  print ('Selected', substrate)
25  # the name of the file where the data will be saved
26  output_filename = 'data_distance_map.dat'
27
28  # select the protein
29  protein = u.select_atoms('protein')
30
31  # create column titles with the format "time, res1, res2, res3"
32  column_titles = ['time(ps)', ]
33  # preselect the heavy atoms for each residue (optimisation)
34  residues = []
35  for i, residue in enumerate(protein.residues, start=1):
```

```
36  column_titles.append(residue.resname + '.' + str(residue.resid) + '.' +
    ↪ str(i))
37  residues.append(residue.atoms.select_atoms('not name H*'))
38
39  assert len(protein.residues) == len(residues), 'There should be heavy atom
    ↪ indices for each residue'
40
41  data = []
42  # adjust the step of the trajectory
43  for ts in u.trajectory[::1]:
44  row = [ts.time, ]
45  for res_hatoms in residues:
46    # find the minimum distance from the heavy residues to the substrate
47    mindst = np.min(distance_array(res_hatoms.positions, substrate.positions,
      ↪ box=ts.dimensions))
48    row.append(mindst)
49    data.append(row)
50    print ("Done timeframe (ns):", ts.time / 1000)
51
52  # save the data with numpy
53  np.savetxt(output_filename, data, fmt='%.2f', header='
    ↪ '.join(column_titles))
```

The distances from the residues to the surface are saved as a 2D matrix, making it easy to plot the results in the form of a heatmap as previously shown in Figure 3.7, right or Figure 5.4. The following plotting script relies on the matplotlib python library.

```
1
2     #!/usr/bin/env python3
3     # -*- coding: utf-8 -*-
```

```
4       """
5       Visualise the distance map representing residues-surface distances
        ↪ calculated using the
6       accompanying python script.
7
8       Please cite the following publication if you find this script useful:
9       Bieniek, M. K. et al. (2019) 'Minor Chemistry Changes Alter Surface
        ↪ Hydration to Control Fibronectin Adsorption and Assembly into
        ↪ Nanofibrils', Advanced Theory and Simulations. John Wiley & Sons, Ltd,
        ↪ p. 1900169. doi: 10.1002/adts.201900169.
10      """
11
12      import matplotlib
13      import matplotlib.pyplot as plt
14      import numpy as np
15      from matplotlib import cm
16
17
18      # paths which have to be adjusted by the user
19      data_filepath = "data_distance_map.dat"
20      plot_output_filepath = "distance_map.png"
21
22      # general configuration for matplotlib
23      matplotlib.rcParams.update({'font.size': 8})
24      plt.figure(figsize=(6, 2))
25      # set the style -
        ↪ https://matplotlib.org/3.1.1/gallery/style_sheets/style_sheets_reference.html
26      #plt.style.use("ggplot")
27      # one plot in x dimension, one in y, and select the first plot
```

```
28      plt.subplot(1, 1, 1)
29      plt.title('EA10 Title')
30      plt.xlabel('Time (ns)')
31      plt.ylabel('Residues')
32
33      # read the residue names first
34      # this is format dependant and depends on how it was saved in the first
        ↪  place
35      resnames = open(data_filepath).readline().split('time(ps)')[1].split()
36      resnames = [n.split('.')[0] + ' ' + n.split('.')[1] for n in resnames]
37
38      # load the data with numpy
39      # plot only every 10th frame (every 10th row), in my case that's a step
        ↪  of 1 ns
40      data = np.loadtxt(data_filepath, comments='#')
41      assert len(resnames) == data.shape[1] - 1, \
42      'The number of residues does not correspond to the number of data columns
        ↪  ' \
43      '(no of residues + 1 column for time)'
44      # remove the time column
45      heatmap = data[:, list(range(1, len(resnames) + 1))]
46      # the x axis should be time, and the y axis should be residues
47      heatmap = heatmap.T
48      # plot while ignoring distances above 20 angstroms
49      # remember to adjust colour map: jet_r is jet is reversed
50      # list of colour maps:
        ↪  https://matplotlib.org/3.1.0/tutorials/colors/colormaps.html
51      contactmap = plt.pcolormesh(heatmap, cmap=cm.jet_r, vmin=0, vmax=20)
52
```

```
53      # prepare places for the yticks (residue names)
54      tick_indices = np.array(list(range(0, len(resnames) + 1, 20)))
55      # shift each label to apply in the middle of its "square"
56      # adjust the fontsize of the y ticks
57      plt.yticks(tick_indices + 0.5, [resnames[i] for i in tick_indices],
        ↪  fontsize=6)
58
59      # create the legend bar in the figure, the ticks correspond to the
        ↪  distances
60      cbar = plt.colorbar(contactmap, ticks=[0, 5, 10, 15, 20], fraction=0.02,
        ↪  pad=0.04)
61      cbar.ax.set_yticklabels(['0$\\rm \AA$', '5$\\rm \AA$', '10$\\rm \AA$',
        ↪  '15$\\rm \AA$', '>20$\\rm \AA$'])
62
63      plt.tight_layout()
64      plt.savefig(plot_output_filepath, dpi=300)
65      # plt.show()
```

Spatial Density Map

The hydration of the surface was analysed using a Spatial Density Map (SDM). In order to obtain an SDM, for each of the constituent molecules of the self-assembled monolayers I extracted the local environment comprising water molecules and the functional groups. Then, I superimposed the functional groups onto each other and rearranged the water molecules to reflect their correct position with respect to their original functional group. The resulting SDM can provide further visual insights into the patterns of behaviour of the surroundings.

In the following script, the SDM of water molecules and the functional groups are extracted and superimposed accordingly.

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
"""
In order to understand the environment composition / density with respect
to the reference molecule, superimpose a selected structure and reorient
 the
environment accordingly. Save the superimposed structures with the
surrounding water molecules in a .pdb file.

Instructions: read carefully the script and update the configuration
 accordingly.

Thanks to:
MDAnalysis: https://www.mdanalysis.org/
This hydration analysis was inspired by ANGULA created by Luis Carlos
 Pardo:
https://gcm.upc.edu/en/members/luis-carlos/angula/ANGULA
"""
import MDAnalysis as mda
from MDAnalysis.analysis.align import rotation_matrix
import matplotlib.pyplot as plt


def fix_pbc(coord, pbc):
    """
    Move the coordinate to the right PBC.
    Assumes that the translation to the origin has been done.
    And that the selection is smaller than half the PBC.
    :param coord: coordinate along the pbc axis
```

```
27      :param pbc: pbc size
28      """
29      if coord > pbc / 2:
30        return coord - pbc
31      elif coord < -(pbc / 2):
32        return coord + pbc
33
34      return coord
35
36  # load the simulation in MDAnalysis
37  u = mda.Universe('sam_hydrated.gro', 'sam_hydrated_step10ns.xtc')
38  # the example input file is a Self-assembled Monolayer (SAM) with 252
    ↪  residues namd PEAC
39
40  # use the first frame of our trajectory as the reference structure
41  # which will be used to superimpose other frames
42  # select four atoms which include an ester as our reference
43  ea_template = u.select_atoms('resname PEAC and name C19 O2 O1 C20 and resid
    ↪  1', updating=False)
44  # shift the origin to be the position of the C20 atom
45  ea_template = ea_template.atoms.translate(-ea_template.select_atoms('name
    ↪  C20')[0].position)
46  ea_ref_pos = ea_template.positions
47
48  # define the output .pdb structure where the superimposed frames will be
    ↪  stored. Note: it has to be a .pdb
49  # due to the flexible number of water molecules found.
50  output = 'sdm.pdb'
51
```

```
52  with mda.Writer(output) as W:
53    rmsds = []
54    # for each frame
55    for ts in u.trajectory:
56      print("Time (ns):", ts.time / 1000)
57      # for each molecule in the self-assembled monolayer
58      for res in u.select_atoms('resname PEAC').residues:
59        # select the residue together with the water oxygens within 5.8 of
          ↪ the C20 atom in the residue
60        sel = u.select_atoms('resid %d or (name OW and around 5.8 (resid %d
          ↪ and name C20))'
61        % (res.resid, res.resid)).residues.atoms
62        original_positions = sel.positions
63        structural_template = sel.select_atoms('name C19 O2 O1 C20')
64
65        # set the position of c20 to be the origin
66        c20 = structural_template.select_atoms('name C20')[0]
67        sel.translate(-c20.position)
68
69        # correct the PBC
70        for atom in sel:
71          corrected_x = fix_pbc(atom.position[0], u.dimensions[0])
72          corrected_y = fix_pbc(atom.position[1], u.dimensions[1])
73          corrected_z = fix_pbc(atom.position[2], u.dimensions[2])
74
75          atom.position = (corrected_x, corrected_y, corrected_z)
76
77        # find the rotation matrix necessary to superimpose the structure
          ↪ against the reference
```

```
78          R, rmsd = rotation_matrix(structural_template.atoms.positions,
            ↪  ea_ref_pos)
79          # monitor the quality of the superimposition
80          rmsds.append(rmsd)
81          # apply the rotation matrix to the molecule, including the
            ↪  surrounding water
82          sel.rotate(R)
83          # save the superimposed atoms in a .pdb format
84          W.write(sel)
85
86
87    plt.title('Superimposed against reference')
88    plt.ylabel('RMSD')
89    plt.hist(rmsds)
90    plt.show()
```

SDM Post Processing After the superimposition and extraction of the surrounding water molecules, one has to extract the specific atoms out of which an SDM can be compiled. In this case, the oxygen atoms from the water molecules are chosen. The output from the previous script has a PDB file format. In order to extract only the lines describing the oxygen atoms, one can select lines which contain the two letters OW. An example using the command *grep* looks like follows "grep OW sdm.pdb > sdm_ow.pdb".

After extracting the positions of the oxygen atoms into the sdm_ow.pdb file, they can be loaded into VMD to generate the graphics. Once loaded into VMD, select from the menu Analysis->Volmap to generate a sdm.dx map (selection=all).

Then, load a single functional group which represents the superimposed molecules, and load the sdm.dx file. Change the visualisation in Graphics->Representations for the sdm.dx to Isosurface and set Draw to Solid Surface. Modify the isovalue to visualise different densities.

Clustering

Clustering with the DBSCAN algorithm was used in order to extract the different conformation of the two molecules with respect to each other. The code used in this process is, similarly, commented and described in this subsection. This technique was used in chapters 7 and 8. Here, the focus is put on the clustering that was used to understand the different FnIII 9-10 domain conformations in chapter 7.

First, the minimum distance between each residue and each other residue is calculated (only heavy atoms are used). In this case, the last frame is used for each system. For each system, the output is a single 2D map of residue-residue distances with diagonal values equal to 0. In this example, 24 different 2D maps are produced, with the systems numbered 0, 15, ..., 345.

```
#!/usr/bin/env python
"""
Find the minimum distance from each residue to each other residue (heavy
↪  atoms only) for all the systems.

In this example, use only the last trajectory frame, which should output a
↪  single 2D n x n matrix,
where n is the number of residues in the protein.
"""

import MDAnalysis as mda
import numpy as np
from MDAnalysis.analysis.distances import distance_array
from collections import OrderedDict
from multiprocessing import Pool


def resres_dsts(system):
```

```
17  directory = str(system) + '/'
18  # open the MD trajectory, adjust the filenames
19  u = mda.Universe(directory + 'npt100ns_centCA_pbcRes_compact.gro',
20  directory + 'npt100ns_centCA_pbcRes_compact_step100ps.xtc')
21
22  # ignore the hydrogen atoms
23  protein_noh = u.select_atoms('protein and not name H*')
24  protein = u.select_atoms('protein')
25
26  # optimisation:
27  # extract which atoms belong to which residues
28  # this is because the distances will be obtained for all atoms
29  resids_atoms = OrderedDict((id, []) for id in set(protein_noh.resids))
30  # the first atom id is 1, but we want to use it as an index so -1
31  [resids_atoms[atom.resid].append(atom.id - 1) for atom in protein_noh.atoms]
32
33  frame = []
34  for ts in u.trajectory[-1:]:
35    # contains distances from all atoms to all other atoms
36    resres_dsts = distance_array(protein.positions, protein.positions,
      ↪  box=ts.dimensions)
37
38    # extract the smallest distances from the matrix
39    # each residue (r1, r2, ..., rn) has a set of heavy atoms,
40    # for the last frame, compute a n-row long distance matrix, with 3
      ↪  residues looking like this:
41    # d1 d2 d3
42    # d1 d2 d3
43    # d1 d2 d3
```

```
44      # d1 d2 d3
45      # ...
46
47      # hardcoded residue IDs
48      for res_id1 in range(1327, 1509 + 1):
49        res1_heavy_atoms = resids_atoms[res_id1]
50        row = []
51        for res_id2 in range(1327, 1509 + 1):
52          res2_heavy_atoms = resids_atoms[res_id2]
53
54          resres_min = np.min(resres_dsts[res1_heavy_atoms][:,
            ↪  res2_heavy_atoms])
55          row.append(resres_min)
56        frame.append(row)
57
58      # save the data
59      columns_titles = 'time(ns) dst_resid1 dst_resid2 ... residn'
60      np.savetxt('output/dstmap_resres_r%d.xvg' % system, frame, fmt='%.3f',
        ↪  header=' '.join(columns_titles))
61
62    # list all the simulations/twists 0, 15, .., 345
63    systems = range(0, 360, 15)
64    # create a Pool of processes for each CPU (16 in my case)
65    pool = Pool(16)
66    # apply the function to each of the simulations
67    pool.map(resres_dsts, systems)
```

Once the residue-residue distances are calculated, producing 24 reside-residue distance maps, the next step is to compare the different 2D maps. In this example, different 2D maps are compared by calculating their similarity S_{ij}:

$S_{ij} = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(m_{i_k} - m_{j_k})^2}$

where m_i and m_j refer to the maps i and j, and n is the number of residues. The value of S_{ij} equal to 0 means that the residue-residue maps i and j are identical. In the following script, the S_{ij} is calculated for each of the residue maps i and j.

```
#!/usr/bin/env python
"""
Create a comparison matrix S_ij which compares different residue-residue maps.
"""
import numpy as np

dstdst_maps = []
for sim in range(0, 360, 15):
    # load the distance-distance map
    resres_maps = np.loadtxt('dstmaps/dstmap_resres_r%d.xvg' % sim)
    # reshape into 1D array
    dstdst_maps.append(resres_maps.reshape(-1))

comparison_matrix = np.zeros((len(dstdst_maps), len(dstdst_maps)))
for i, i_map in enumerate(dstdst_maps[:-1]):
    for j, j_map in enumerate(dstdst_maps[i + 1:], start=i + 1):
        # Compute the root mean square (RMS) between the residue-residue distance maps
        weight = np.sum(np.power(i_map - j_map, 2))
        # normalise
        assert len(i_map) == len(j_map)
        weight /= len(i_map)
        weight = np.sqrt(weight)
```

```
23
24    # the comparison matrix is symmetrical
25    comparison_matrix[j][i] = comparison_matrix[i][j] = weight
26
27    np.savetxt('maps_comparison_matrix.dat', comparison_matrix)
28
```

The output of the previous script is a comparison matrix which can be used to plot the similarity between the different residue-residue maps. This is done in the following matplotlib script.

```
1     #!/usr/bin/env python
2     """
3     Plot a comparison matrix using matplotlib.
4
5     Inspired by
6     https://stackoverflow.com/questions/33282368
7     """
8     import numpy as np
9     import matplotlib
10    import matplotlib.pyplot as plt
11
12    sparse_m = np.loadtxt('maps_comparison_matrix.dat')
13
14    # prepare the mask
15    mask = np.zeros_like(sparse_m)
16    # select the value you want to keep
17    mask[np.triu_indices_from(mask)] = True
18    masked_data = np.ma.array(sparse_m[::-1], mask=mask[::-1])
19
20    plt.figure(figsize=(6, 4))
```

```
21  plt.style.use('ggplot')
22  matplotlib.rcParams.update({'font.size': 10})
23
24  # adjust the cmap and other details
25  im = plt.pcolor(masked_data, cmap='bwr') #, vmin=0, vmax=15)
26  # create and configure the colour bar
27  clb = plt.colorbar(im)
28  # clb.set_ticks([1,2,3])
29  # clb.set_ticklabels([1,2,3])
30
31  plt.xticks(rotation='vertical')
32  plt.xlabel('System r(#)')
33  plt.ylabel('System r(#)')
34
35  # rename the ticks to indicate the system numbers
36  plt.xticks(np.linspace(0.5,23.5, 24), range(0, 360, 15))
37  plt.yticks(np.linspace(0.5,23.5, 24), range(0, 360, 15)[::-1])
38
39  plt.title('Residue-Residue Map Similarity')
40
41  plt.tight_layout()
42  plt.savefig('S_ij_matrix.png', dpi=300)
43  plt.show()
44
```

Finally, the clustering is carried out on the comparison matrix in order to group similar FnIII 9-10 conformation. For this, the DBSCAN was chosen and its use is presented in the following script.

```
1  #!/usr/bin/env python
```

```
"""
Apply DBSCAN clustering to the comparison matrix in
order to extract the different possible states
existing in the sampled interdomain space.

For the official documentation of DBSCAN, please see
https://scikit-learn.org/
"""
import numpy as np
from sklearn.cluster import DBSCAN

cmp_matrix = np.sqrt(np.loadtxt("maps_comparison_matrix.dat"))

# EPS: The maximum distance between two samples for one to be considered
↪ as in the neighborhood of the other.
# set the EPS value to 1.5 Angstrom which due to the 2D matrix
↪ representation of any two pictures
# translates to 0.75 Angstrom. Ie two different protein conformations have
↪ to have RMSD of less than 0.75 Angstrom
# to be considered as in the neighbourhood of each other.
eps = 1.5
db = DBSCAN(metric="precomputed", eps=eps, min_samples=1).fit(X=cmp_matrix)

# extract the clusters from the field
labels = db.labels_
# count the unique labels
num_clusters = len(set(labels)) - (1 if -1 in labels else 0)
print('Estimated number of clusters: %d' % num_clusters)

```

```
28    # count the number of systems classified as noise
29    num_noise = list(labels).count(-1)
30    print('Estimated number of noise points: %d' % num_noise)
31
32    # print the found clusters
33    for label in set(labels):
34      # extract the indices of the systems with the label
35      indices = np.where(labels == label)
36      # multily each index by 15 to recover the system's name (initial degree
        ↪  tilt)
37      print('Label: %d,' % label + ' Systems:', indices[0] * 15)
```

The code will be made available online in the form of a tutorial under `https://github.com/bieniekmateusz/publications`.

A.2 MDAnalysis

In this section I focus on the work carried out on the open source python software MDAnalysis, which was carried out together with Paul Smith, a fellow PhD student at King's College London. This work began with a small patch submitted to repair a bug (`https://github.com/MDAnalysis/mdanalysis/pull/1759`) due to which the parameter "start time" in the function SurvivalProbability was ignored. This was followed by substantial reimplementation of the package MDAnalysis.analysis.waterdynamics along with the corresponding unit tests (`https://github.com/MDAnalysis/mdanalysis/pull/1995` and `https://github.com/MDAnalysis/mdanalysis/pull/2226`).

The survival probability (SP) describes the propensity of selected molecules to remain in a selected environment. For example, assume that you are trying to understand the hydration of a specific chemical group in a drug. SP tells us how long on average the water molecules survives within a certain distance. The example output includes two series: the timeseries 0, 1,

2, 3, ..., x picoseconds, and the corresponding values for each time points specifying how likely the water molecules are to be found in the selected region: 100%, 50%, 25%, 12%, ..., 0%. This concept can also be referred to as discrete autocorrelation function.

A closely related problem to the discrete autocorrelation function is the problem of intermittency. In order to understand the survival of oxygen atoms in the water molecules it should be noted that some of the molecules leave the selection for a very short fraction of time only to return and remain in the area of interest.

The previous implementation of intermittency was naive, consequently overestimating the survival probability. Specifically, if an atom was was trapped on the border of the selected region, it might leave and return regularly. In the previous implementation of the intermittency, the time in which the atom was not found in the region of interest was summed together, and checked against the user's requirements. For example, if the user specified that an atom can be absent for 5 ps in a window of 20 ps, and the atom was absent 6 times where each absence lasted 1 ps, then the atom was classified as absent in that frame. However, the atom's absence lasting the frame's 5 ps period 15 - 20 ps would be quite different to the same atoms leaving for 1 ps and returning each time. For this reason, we defined and implemented consecutive intermittency, as described below.

The consecutive intermittency allows the user to define for how long at any time an atom is allowed to leave and return. This way, any trapped atom on the border of the selected region is not disqualified as absent. Furthermore, due to the discrete nature of MD trajectory analysis, the consecutive intermittency approach is a more suitable and clearer approach.

In addition to the refinement of intermittency we implemented the option to define "step" to sample the simulation. Such a step defines how many frames are skipped between two data points. We made it possible to use the step in the analysis together with the intermittency, ensuring that the two are consistent with each other. This can be particularly useful when working with very long simulations which do not need to use every recorded frame.

The discrete autocorrelation and the intermittency functions were factored out of the SurvivalProbability and are now accessible as separate tools for other analysis within the MDAnalysis

software. This means that these tools are now easier to use internally as well as externally. With these well defined tools it is now possible to replace other implementations of autocorrelation in MDAnalysis, such as the analysis of hydrogen bond autocorrelation (work in progress).

As part of the reimplementation we redefined the tests to validate the correctness of the code. This encompasses a well defined set of tests which check that the functions perform what they set out to do.

Below I present the documented code for the autocorrelation and the intermittency function which reside in the file autocorrelation.py which itself can be found under the package MDAnalysis.analysis.utils. This is currently the default implementation of SurvivalProbability (MDAnalysis.analysis.waterdynamics) in the MDAnalysis software. In addition, the testing environment ensuring the correctness of the implementation is presented.

```
1  # -*- Mode: python; tab-width: 4; indent-tabs-mode:nil; coding:utf-8 -*-
2  # vim: tabstop=4 expandtab shiftwidth=4 softtabstop=4
3  #
4  # MDAnalysis --- https://www.mdanalysis.org
5  # Copyright (c) 2006-2017 The MDAnalysis Development Team and contributors
6  # (see the file AUTHORS for the full list of names)
7  #
8  # Released under the GNU Public Licence, v2 or any higher version
9  #
10 # Please cite your use of MDAnalysis in published work:
11 #
12 # R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L.
   ↪ Seyler,
13 # D. L. Dotson, J. Domanski, S. Buchoux, I. M. Kenney, and O. Beckstein.
14 # MDAnalysis: A Python package for the rapid analysis of molecular
   ↪ dynamics
```

```
15  # simulations. In S. Benthall and S. Rostrup editors, Proceedings of the
    ↪ 15th
16  # Python in Science Conference, pages 102-109, Austin, TX, 2016. SciPy.
17  # doi: 10.25080/majora-629e541a-00e
18  #
19  # N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein.
20  # MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics
    ↪ Simulations.
21  # J. Comput. Chem. 32 (2011), 2319--2327, doi:10.1002/jcc.21787
22  #
23
24  import numpy as np
25  from copy import deepcopy
26
27
28  def autocorrelation(list_of_sets, tau_max, window_step=1):
29    r"""The descrete implementation of the autocorrelation function.
30
31    Parameters
32    ----------
33    list_of_sets : list
34      List of sets
35    tau_max : int
36      The last tau (inclusive) for which to carry out autocorrelation.
37    window_step : int, optional
38      The step for the t0 to perform autocorrelation (without the overlap).
  ↪ Default is 1..
39
40    Returns
```

```
    --------
    tau_timeseries : list of int
      the tau for which the autocorrelation was calculated
    timeseries : list of int
      the autocorelation values for each of the tau values
    timeseries_data : list of list of int
      the raw data from which the autocorrelation is computed. The time
    ↪ dependant evolution can be investigated.

    .. versionadded:: 0.19.2
    """
    tau_timeseries = list(range(1, tau_max + 1))
    timeseries_data = [[] for _ in range(tau_max)]

    # calculate autocorrelation
    for t in range(0, len(list_of_sets), window_step):
      Nt = len(list_of_sets[t])

      if Nt == 0:
        continue

      # check the current window
      for tau in tau_timeseries:
        if t + tau >= len(list_of_sets):
          break

        # IDs that survive from t to t + tau and at every frame in between
        Ntau = len(set.intersection(*list_of_sets[t:t + tau + 1]))
        timeseries_data[tau - 1].append(Ntau / float(Nt))
```

```
69
70      timeseries = [np.mean(x) for x in timeseries_data]
71
72      # at time 0 the value has to be one
73      tau_timeseries.insert(0, 0)
74      timeseries.insert(0, 1)
75
76      return tau_timeseries, timeseries, timeseries_data
77
78
79    def correct_intermittency(list_of_sets, intermittency):
80      """
81      Pre-process Consecutive Intermittency with a single pass over the data.
82      If an atom is absent for a number of frames equal or smaller
83      than the parameter intermittency, update the data and remove the
    ↪ absence(s).
84      For example, having the sequence [7,A,A,7], where A=absence and the
    ↪ digit represents
85      an atom with ID=7, setting the intermittency=2 will results in the
    ↪ updated sequence [7,7,7,7].
86
87      Parameters
88      ----------
89      id_list: list of sets
90        returns a new list with added IDs which disappeared for <= :param
    ↪ intermittency
91      intermittency: int
92        the max gap allowed which will be corrected
93      """
```

```
94
95      if intermittency == 0:
96        return list_of_sets
97
98      # copy the entire dataset
99      list_of_sets = deepcopy(list_of_sets)
100
101     for i, ids in enumerate(list_of_sets):
102       # initially update each ID as seen 0 frames ago (now)
103       seen_frames_ago = {i: 0 for i in ids}z
104       for j in range(1, intermittency + 2):
105         for atomid in seen_frames_ago.keys():
106           # no more frames to check
107           if i + j >= len(list_of_sets):
108             continue
109
110           # if the atom is absent, record it
111           if not atomid in list_of_sets[i + j]:
112             # increase its absence counter
113             seen_frames_ago[atomid] += 1
114             continue
115
116           # the atom was present in the last frame
117           if seen_frames_ago[atomid] == 0:
118             continue
119
120           # it was absent more times than allowed
121           if seen_frames_ago[atomid] > intermittency:
122             continue
```

```
123
124             # the atom was absent but returned (within <= intermittency_value)
125             # add it to the frames where it was absent.
126             # ie. Introduce the corrections.
127             for k in range(seen_frames_ago[atomid], 0, -1):
128             list_of_sets[i + j - k].add(atomid)
129
130         seen_frames_ago[atomid] = 0
131     return list_of_sets
132
```

Below I present the class SurvivalProbability which makes use of the updated autocorrelation and intermittency functions. For further details, please see the comments as well as the online documentation which can be found at `https://www.mdanalysis.org/docs/documentation_pages/analysis/waterdynamics.html`.

```
1
2   class SurvivalProbability(object):
3   r"""
4   Survival Probability (SP) gives the probability for a group of particles
    ↪  to remain in a certain region.
5   The SP is given by:
6
7   .. math::
8   P(\tau) = \frac1T \sum_{t=1}^T \frac{N(t,t+\tau)}{N(t)}
9
10  where :math:`T` is the maximum time of simulation, :math:`\tau` is the
11  timestep, :math:`N(t)` the number of particles at time :math:`t`, and
12  :math:`N(t, t+\tau)` is the number of particles at every frame from
    ↪  :math:`t` to `\tau`.
```

```
13
14
15    Parameters
16    ----------
17    universe : Universe
18      Universe object
19    selection : str
20      Selection string; any selection is allowed. With this selection you
21      define the region/zone where to analyze, e.g.: "resname SOL and around 5
      ↪  (resid 10)". See `SP-examples`_.
22    verbose : Boolean, optional
23      When True, prints progress and comments to the console.
24
25
26    .. versionadded:: 0.11.0
27
28    """
29
30    def __init__(self, universe, selection, t0=None, tf=None, dtmax=None,
       ↪  verbose=False):
31      self.universe = universe
32      self.selection = selection
33      self.verbose = verbose
34
35      # backward compatibility
36      self.start = self.stop = self.tau_max = None
37      if t0 is not None:
38        self.start = t0
```

```
39        warnings.warn("t0 is deprecated, use run(start=t0) instead",
          ↪  category=DeprecationWarning)
40
41      if tf is not None:
42        self.stop = tf
43        warnings.warn("tf is deprecated, use run(stop=tf) instead",
          ↪  category=DeprecationWarning)
44
45      if dtmax is not None:
46        self.tau_max = dtmax
47        warnings.warn("dtmax is deprecated, use run(tau_max=dtmax) instead",
          ↪  category=DeprecationWarning)
48
49
50    def run(self, tau_max=20, start=0, stop=None, step=1, residues=False,
      ↪  intermittency=0, verbose=False):
51      """
52      Computes and returns the Survival Probability (SP) timeseries
53
54      Parameters
55      ----------
56      start : int, optional
57        Zero-based index of the first frame to be analysed
58      stop : int, optional
59        Zero-based index of the last frame to be analysed (inclusive)
60      step : int, optional
61        Jump every `step`-th frame. This is compatible but independant of the
    ↪  taus used, and it is good to consider
62        using the  `step` equal to `tau_max` to remove the overlap.
```

```
63        Note that `step` and `tau_max` work consistently with intermittency.
64    tau_max : int, optional
65        Survival probability is calculated for the range 1 <= `tau` <=
    ↪ `tau_max`
66    residues : Boolean, optional
67        If true, the analysis will be carried out on the residues (.resids)
    ↪ rather than on atom (.ids).
68        A single atom is sufficient to classify the residue as within the
    ↪ distance.
69    intermittency : int, optional
70        The maximum number of consecutive frames for which an atom can leave
    ↪ but be counted as present if it returns
71        at the next frame. An intermittency of `0` is equivalent to a
    ↪ continuous survival probability, which does
72        not allow for the leaving and returning of atoms. For example, for
    ↪ `intermittency=2`, any given atom may
73        leave a region of interest for up to two consecutive frames yet be
    ↪ treated as being present at all frames.
74        The default is continuous (0).
75    verbose : Boolean, optional
76        Print the progress to the console
77
78    Returns
79    -------
80    tau_timeseries : list
81        tau from 1 to `tau_max`. Saved in the field tau_timeseries.
82    sp_timeseries : list
83        survival probability for each value of `tau`. Saved in the field
    ↪ sp_timeseries.
```

```
84      sp_timeseries_data: list
85        raw datapoints from which the average is taken (sp_timeseries).
86        Time dependancy and distribution can be extracted.
87      """
88
89      # backward compatibility (and priority)
90      start = self.start if self.start is not None else start
91      stop = self.stop if self.stop is not None else stop
92      tau_max = self.tau_max if self.tau_max is not None else tau_max
93
94      # sanity checks
95      if stop is not None and stop >= len(self.universe.trajectory):
96        raise ValueError("\"stop\" must be smaller than the number of frames in
          ↪  the trajectory.")
97
98      if stop is None:
99        stop = len(self.universe.trajectory)
100     else:
101       stop = stop + 1
102
103     if tau_max > (stop - start):
104       raise ValueError("Too few frames selected for given tau_max.")
105
106     # preload the frames (atom IDs) to a list of sets
107     self._selected_ids = []
108
109     # Improve - to parallise: the section should be rewritten so that this
        ↪  loop only creates a list of indices,
110     # on which the parallel _single_frame can be applied.
```

```
111
112         # skip frames that will not be used
113         # Example: step 5 and tau 2: LLLSS LLLSS, ... where L = Load, and S =
            ↪ Skip
114         # Intermittency means that we have to load the extra frames to know if
            ↪ the atom is actually missing.
115         # Say step=5 and tau=1, intermittency=0: LLSSS LLSSS
116         # Say step=5 and tau=1, intermittency=1: LLLSL LLLSL
117         frame_loaded_counter = 0
118         # only for the first window (frames before t are not used)
119         frames_per_window = tau_max + 1 + intermittency
120         # This number will apply after the first windows was loaded
121         frames_per_window_subsequent = (tau_max + 1) + (2 * intermittency)
122         num_frames_to_skip = max(step - frames_per_window_subsequent, 0)
123
124         frame_no = start
125         while frame_no < stop:      # we have already added 1 to stop, therefore
            ↪ <
126           if num_frames_to_skip != 0 and frame_loaded_counter ==
              ↪ frames_per_window:
127             logger.info("Skipping the next %d frames:" % num_frames_to_skip)
128             frame_no += num_frames_to_skip
129             frame_loaded_counter = 0
130             # Correct the number of frames to be loaded after the first window
                ↪ (which starts at t=0, and
131             # intermittency does not apply to the frames before)
132             frames_per_window = frames_per_window_subsequent
133             continue
134
```

```
135         # update the frame number
136         self.universe.trajectory[frame_no]
137 
138         logging.info("Loading frame:", self.universe.trajectory.ts)
139         atoms = self.universe.select_atoms(self.selection)
140 
141         # SP of residues or of atoms
142         ids = atoms.residues.resids if residues else atoms.ids
143         self._selected_ids.append(set(ids))
144 
145         frame_no += 1
146         frame_loaded_counter += 1
147 
148     # adjust for the frames that were not loaded (step>tau_max + 1),
149     # and for extra frames that were loaded (intermittency)
150     window_jump = step - num_frames_to_skip
151 
152     self._intermittent_selected_ids =
        ↪  correct_intermittency(self._selected_ids, intermittency=intermittency)
153     tau_timeseries, sp_timeseries, sp_timeseries_data =
        ↪  autocorrelation(self._intermittent_selected_ids,
154     tau_max, window_jump)
155 
156     # warn the user if the NaN are found
157     if all(np.isnan(sp_timeseries[1:])):
158         logging.warning('NaN Error: Most likely data was not found. Check your
            ↪  atom selections. ')
159 
160     # user can investigate the distribution and sample size
```

```
161     self.sp_timeseries_data = sp_timeseries_data
162
163     self.tau_timeseries = tau_timeseries
164     self.sp_timeseries = sp_timeseries
165     return self
```

The abstracted implementation of the autocorrelation and intermittency is thoroughly tested using test cases, which are described below.

```
1   def test_autocorrelation_alwaysPresent():
2     input = [{1, 2}, {1, 2}, {1, 2}, {1, 2}, {1, 2}, {1, 2}, {1, 2}]
3     tau_timeseries, sp_timeseries, sp_timeseries_data = autocorrelation(input,
      ↪  tau_max=3)
4     assert all(np.equal(sp_timeseries, 1))
5
6   def test_autocorrelation_perfTest():
7     # generate a list of sets
8     import random
9     input = [{x for x in range(50)} for x in range(1000 * 1000)]
10    import time
11    st = time.time()
12    #input = [{1, 2}, {1, 2}, {1, 2}, {1, 2}, {1, 2}, {1, 2}, {1, 2}]
13    tau_timeseries, sp_timeseries, sp_timeseries_data = autocorrelation(input,
      ↪  tau_max=3)
14    print('Comuptation Time', time.time() - st)
15    assert False
16
17
18  def test_autocorrelation_definedTaus():
```

```
19      input_ids = [{9, 8, 7}, {8, 7, 6}, {7, 6, 5}, {6, 5, 4}, {5, 4, 3}, {4, 3,
        ↪  2}, {3, 2, 1}]
20      tau_timeseries, sp_timeseries, sp_timeseries_data =
        ↪  autocorrelation(input_ids, tau_max=3)
21      assert_almost_equal(sp_timeseries, [1, 2/3., 1/3., 0])
22
23
24    def test_autocorrelation_intermittency1_windowJump_intermittencyAll():
25      """
26      Step leads to skipping frames if (tau_max + 1) + (intermittency * 2) <
    ↪   step.
27      No frames should be skipped so intermittency should be applied to all.
28      """
29      input_ids = [{2, 3}, {3,}, {2, 3}, {3,}, {2,}, {3,}, {2, 3}, {3,}, {2, 3},
        ↪  {2, 3}]
30      corrected = correct_intermittency(input_ids, intermittency=1)
31      tau_timeseries, sp_timeseries, sp_timeseries_data =
        ↪  autocorrelation(corrected, tau_max=2,
32      window_step=5)
33      assert all((x == {2, 3} for x in corrected))
34      assert_almost_equal(sp_timeseries, [1, 1, 1])
35
36
37    def test_autocorrelation_windowBigJump():
38      #The empty sets are ignored (no intermittency)
39      input_ids = [{1}, {1}, {1}, set(), set(), {1}, {1}, {1}, set(), set(),
        ↪  {1}, {1}, {1}]
40      tau_timeseries, sp_timeseries, sp_timeseries_data =
        ↪  autocorrelation(input_ids, tau_max=2, window_step=5)
```

```
41      assert_almost_equal(sp_timeseries, [1, 1, 1])
42
43
44    def test_autocorrelation_windowBigJump_absence():
45      # In the last frame the molecules are absent
46      input_ids = [{1}, {1}, {1}, set(), set(), {1}, {1}, {1}, set(), set(),
        ↪  {1}, set(), set()]
47      tau_timeseries, sp_timeseries, sp_timeseries_data =
        ↪  autocorrelation(input_ids, tau_max=2, window_step=5)
48      assert_almost_equal(sp_timeseries, [1, 2/3., 2/3.])
49
50
51    def test_autocorrelation_intermittency1_many():
52      input_ids = [{1}, set(), {1}, set(), {1}, set(), {1}, set(), {1}, set(),
        ↪  {1}, set(), {1}, set(), {1}]
53      corrected = correct_intermittency(input_ids, intermittency=1)
54      tau_timeseries, sp_timeseries, sp_timeseries_data =
        ↪  autocorrelation(corrected, tau_max=14,
55      window_step=5)
56      assert_almost_equal(sp_timeseries, [1] * 15)
57
58
59    def test_autocorrelation_intermittency2_windowBigJump():
60      # The intermittency corrects the last frame
61      input_ids = [{1}, {1}, {1}, set(), set(), {1}, {1}, {1}, set(), set(),
        ↪  {1}, set(), set(), {1}]
62      corrected = correct_intermittency(input_ids, intermittency=2)
63      tau_timeseries, sp_timeseries, sp_timeseries_data =
        ↪  autocorrelation(corrected, tau_max=2,
```

```
64      window_step=5)
65      assert_almost_equal(sp_timeseries, [1, 1, 1])
66
67
68    def test_intermittency_none():
69      # No changes asked - returns the same data
70      input_ids = [{1}, {1}, {1}, set(), set(), {1}, {1}, {1}, set(), set(),
        ↪  {1}, set(), set(), {1}]
71      corrected = correct_intermittency(input_ids, intermittency=0)
72      assert all(x == y for x,y in zip(input_ids, corrected))
73
74
75    def test_intermittency_1and2():
76      # The maximum gap in the dataset is 2, so the IDs are always present
        ↪  after correction
77      input_ids = [{9, 8}, set(), {8, }, {9, }, {8, }, set(), {9, 8}, set(), {8,
        ↪  }, {9, 8, }]
78      corrected = correct_intermittency(input_ids, intermittency=2)
79      assert all((x == {9, 8} for x in corrected))
80
81
82    def test_intermittency_2tooShort():
83      #The IDs are abscent for too long/
84      input_ids = [{9,}, {}, {}, {}, {9,}, {}, {}, {}, {9,}]
85      corrected = correct_intermittency(input_ids, intermittency=2)
86      assert all(x == y for x, y in zip(input_ids, corrected))
87
88
89    def test_intermittency_setsOfSets():
```

```
90      # Verificaiton for the case of hydrogen bonds (sets of sets)
91      input_ids = [{frozenset({1,2}), frozenset({3, 4})},set(), set(),
92      {frozenset({1, 2}), frozenset({3, 4})}, set(), set(),
93      {frozenset({1, 2}), frozenset({3, 4})}, set(), set(),
94      {frozenset({1, 2}), frozenset({3, 4})}]
95      corrected = correct_intermittency(input_ids, intermittency=2)
96      assert all((x == {frozenset({1, 2}), frozenset({3, 4})} for x in
        ↪  corrected))
97
98
99    def test_SurvivalProbability_intermittency1and2(universe):
100     """
101     Intermittency of 2 means that we still count an atom if it is not
    ↪  present for up to 2 consecutive frames,
102     but then returns at the following step.
103     """
104     with patch.object(universe, 'select_atoms') as select_atoms_mock:
105       ids = [(9, 8), (), (8,), (9,), (8,), (), (9, 8), (), (8,), (9, 8)]
106       select_atoms_mock.side_effect = lambda selection: Mock(ids=ids.pop())
          ↪  # atom IDs fed set by set
107       sp = waterdynamics.SurvivalProbability(universe, "")
108       sp.run(tau_max=3, stop=9, verbose=True, intermittency=2)
109       assert all(x == {9, 8} for x in sp._intermittent_selected_ids)
110       assert_almost_equal(sp.sp_timeseries, [1, 1, 1, 1])
111
112
113   def test_SurvivalProbability_intermittency2lacking(universe):
114     """
115     If an atom is not present for more than 2 consecutive frames,
```

```
116     it is considered to have left the region.
117     """
118     with patch.object(universe, 'select_atoms') as select_atoms_mock:
119       ids = [(9,), (), (), (), (9,), (), (), (), (9,)]
120       select_atoms_mock.side_effect = lambda selection: Mock(ids=ids.pop())
          ↪  # atom IDs fed set by set
121       sp = waterdynamics.SurvivalProbability(universe, "")
122       sp.run(tau_max=3, stop=8, verbose=True, intermittency=2)
123       assert_almost_equal(sp.sp_timeseries, [1, 0, 0, 0])
124
125
126   def test_SurvivalProbability_intermittency1_step5_noSkipping(universe):
127     """
128     Step leads to skipping frames if (tau_max + 1) + (intermittency * 2) <
    ↪  step.
129     No frames should be skipped.
130     """
131     with patch.object(universe, 'select_atoms') as select_atoms_mock:
132       ids = [(2, 3), (3,), (2, 3), (3,), (2,), (3,), (2, 3), (3,), (2, 3), (2,
          ↪  3)]
133       select_atoms_mock.side_effect = lambda selection: Mock(ids=ids.pop())
          ↪  # atom IDs fed set by set
134       sp = waterdynamics.SurvivalProbability(universe, "")
135       sp.run(tau_max=2, stop=9, verbose=True, intermittency=1, step=5)
136       assert all((x == {2, 3} for x in sp._intermittent_selected_ids))
137       assert_almost_equal(sp.sp_timeseries, [1, 1, 1])
138
139
140   def test_SurvivalProbability_intermittency1_step5_Skipping(universe):
```

```
141     """
142     Step leads to skipping frames if (tau_max + 1) * (intermittency * 2) <
    ↪ step.
143     In this case one frame will be skipped per window.
144     """
145     with patch.object(universe, 'select_atoms') as select_atoms_mock:
146       ids = [(1,), (), (1,), (), (1,), (), (1,), (), (1,), (1,)]
147       beforepopsing = len(ids) - 2
148       select_atoms_mock.side_effect = lambda selection: Mock(ids=ids.pop())
        ↪  # atom IDs fed set by set
149       sp = waterdynamics.SurvivalProbability(universe, "")
150       sp.run(tau_max=1, stop=9, verbose=True, intermittency=1, step=5)
151       assert all((x == {1} for x in sp._intermittent_selected_ids))
152       assert len(sp._selected_ids) == beforepopsing
153       assert_almost_equal(sp.sp_timeseries, [1, 1])
154
```

The contributed MDAnalysis code is available on the MDAnalysis github project (`https://github.com/MDAnalysis/mdanalysis`).

A.3 PyMOL

Paul Smith and I were awarded the Warren L. DeLano Memorial PyMOL Open-Source Fellowship. Our accepted project proposal focused on embedding MDAnalysis into the PyMOL software in order to gain access to the vast functionalities available in MDAnalysis.

Our initial objective was to use MDAnalysis in order to load the molecular dynamics trajectory into memory. The main advantage offered by MDAnalysis is the ability to access any of the trajectory frames without the necessity of loading the previous frames. This approach eases the work with molecular dynamics simulations by allowing the user to load only the coordinates

of a single frame. This way, we reduce the time it takes to "load" the trajectory to minimum, as well as decrease RAM-memory footprint. The downside to this approach is that browsing the trajectory, particularly large trajectories, might become less responsive. This is due to the fact that with each change to a new frame, the next frame has to be fetched. However, this is mitigated by solid state drive (SSD) technology which is becoming ubiquitous, and which offers rapid access to any arbitrary location. Due to SSD technology, the trajectories can be viewed without any substantial delay.

The first prototype of this feature has been presented by Paul Smith in a short youtube video (`https://www.youtube.com/watch?v=Rk55tbK2KMQ`). The video presents how a PyMOL-MDAnalysis user can open and start working on trajectory data without the need to wait for it to be loaded.

MDAnalysis also provides access to a growing set of tools and functions for the analysis of molecular dynamics trajectories. In the second phase of the fellowship, we show the viability of bringing the MDAnalysis tools into PyMOL. As a proof-of-concept, we selected a well optimised and familiar analysis tool for root mean square deviation (RMSD, MDAnalysis.analysis.rms.rmsd) to be embedded into PyMOL. The MDAnalysis RMSD function was made accessible from within of PyMOL with a new commands "mda_rmsd". This marriage is made relatively straight forward with the appropriate use of the kwargs in python. The important part of the new functionality is the interactive plotting. MDAnalysis is unable to visualise the results of the analysis, and for that reason, we introduced matplotlib.

The python plotting library matplotlib was added to provide basic interactive plotting features. Matplotlib was chosen due to its wide use in academia for creating publishing-quality graphs and visualisations.

The added PyMOL RMSD function uses matplotlib to visualise the results overtime. The interactive feature is the ability to press on the RMSD plot which will update PyMOL to visualise the requested time point. Furthermore, the user can easily highlight any section of the RMSD and obtain the resulting histogram. A snapshot of the RMSD functionality is presented in Figure A.1.

After the visualisation, the generated data from the RMSD as well as the plots are stored

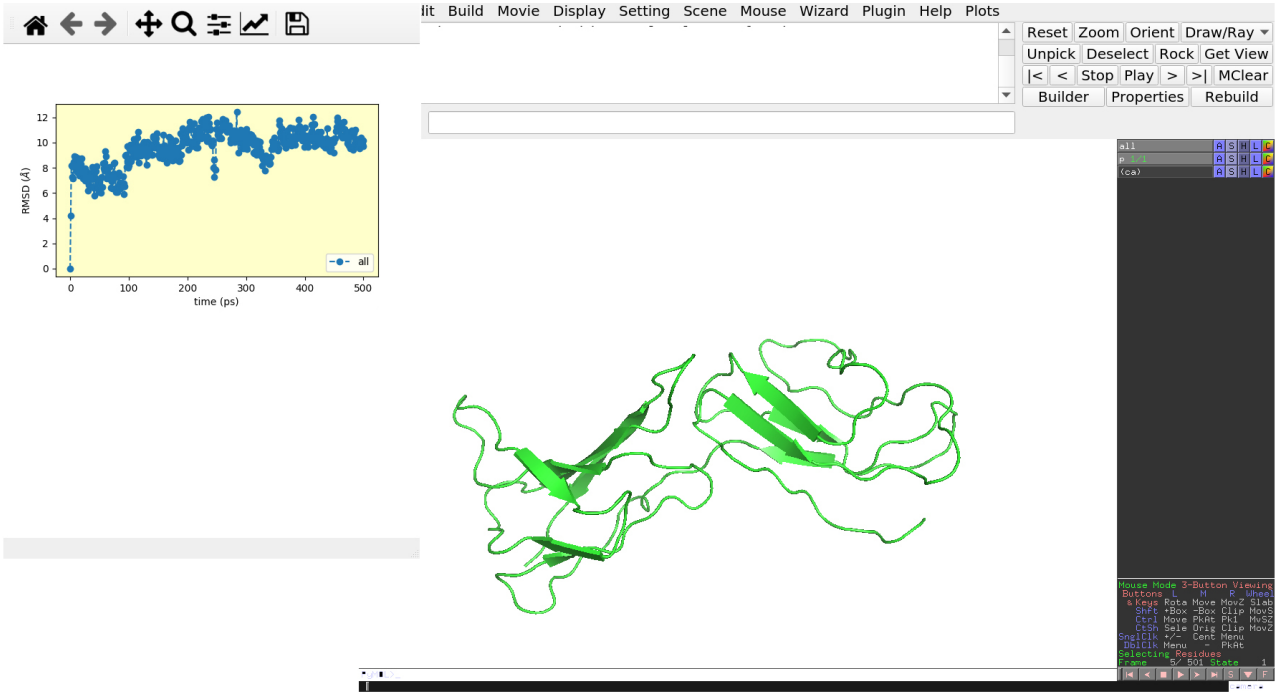


Figure A.1: PyMOL and the RMSD interactive plot.

on the local disk. Therefore the user can easily access any of the previously generated data. Furthermore, the visualising python scripts are stored on the disk along with the data, making it very easy for the user to personalise the script and replot the data.

The RMSD graph can be further modified in a visual manner. This includes features like updating the axis, title of the plot, its location, labels, and others. This is not natively supported by matplotlib and therefore we employed an additional (recently published) python software pylustrator [3]. Pylustrator is compatible with the PyMOL-MDAnalysis marriage: it updates the visualising python scripts, and works with and without PyMOL. An example of an RMSD plot with pylustrator is presented in Figure A.2.

In order to avoid the user having to personalise the overall style each time, we introduced plotting templates. These allow the user to define the style once, before the analysis is carried out. This is made possible by modifying the template plotting files which are used when generating the plots for the first time. These templates reside in the same directory and are easy to share, allowing collaborations and teams to use the same styling.

In addition to the interactive, we added a productivity feature. In our workload, we noticed that frequently the same selections have to be recreated. For example, when working with

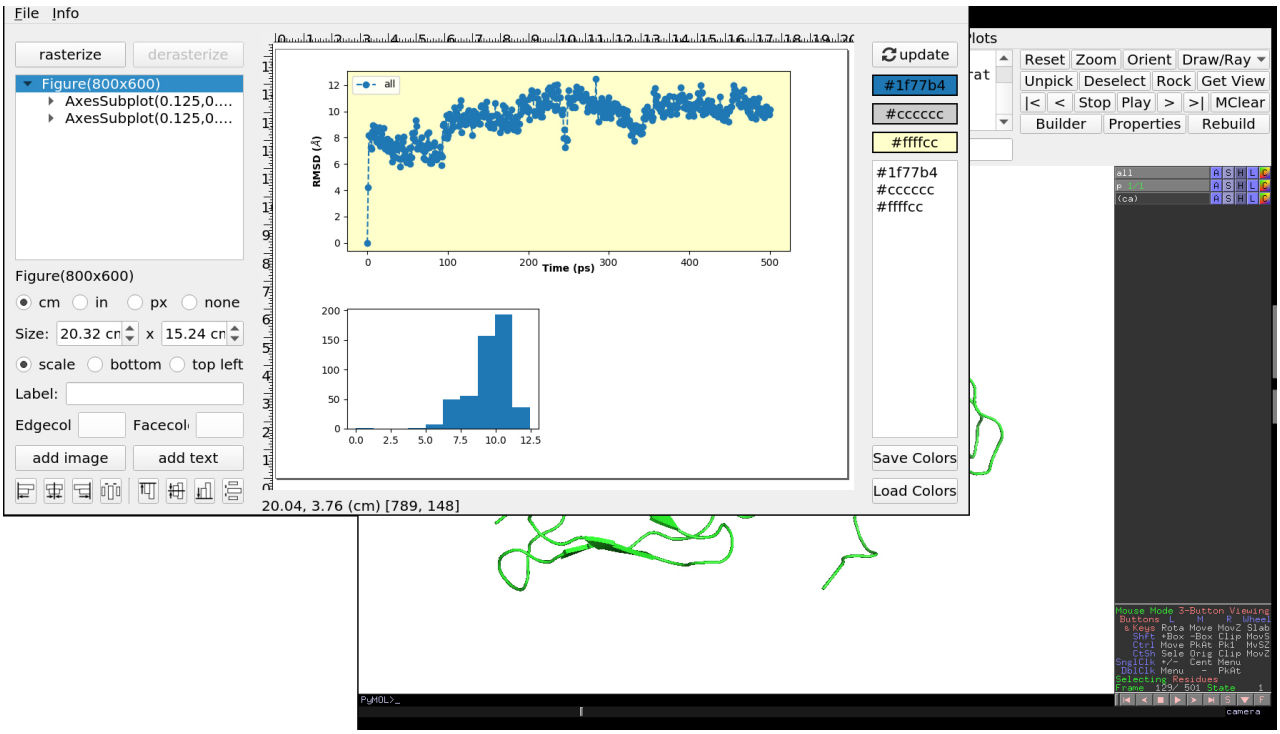


Figure A.2: Pylustrator [3] used in combination with PyMOL-MDAnalysis. Pylustrator was used to bold the labels in the top figure and to decrease the size of the bottom figure. For the documentation of all features please see `https://pylustrator.readthedocs.io/en/latest/`.

one protein, we were continually selecting the same residues that we currently worked on. To increase productivity, we save these selections, and later allow the user to recover them. A screenshot of this feature is presented in Figure A.3

The submitted code together with the work history is available on github (`https://github.com/bieniekmateusz/pymol-open-source/commits/fellows_mp_2018`).

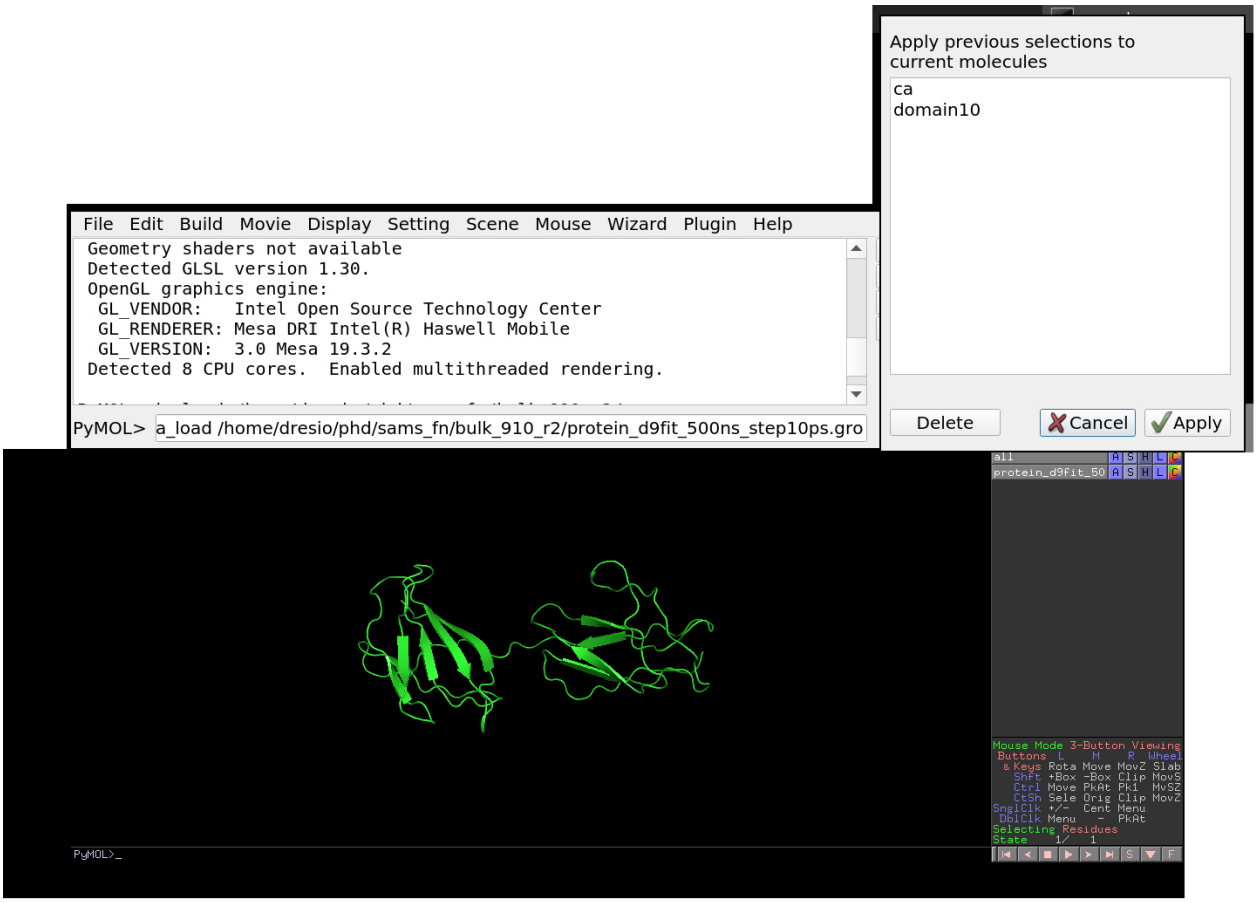


Figure A.3: Reselection feature: when loading a previously used topology or coordinate file, the user is offered the option to recover the previously created selections.